AFIT/GNE/ENP/93M-7

INDENTIFICATION OF SIGNIFICANT OUTLIERS IN TIME SERIES DATA

THESIS

Keri L. Robinson
Captain, USAF

AFIT/GNE/ENP/93M-7

Approved for public release; distribution unlimited

AFIT/GNE/ENP/93M-7

IDENTIFICATION OF SIGNIFICANT OUTLIERS IN TIME SERIES DATA

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Nuclear Engineering

Keri L. Robinson, B.NE

Captain, USAF

March 1993

Preface

The purpose of this study was to develop a methodology for detecting significant outliers in time series data that would either compliment or supplant the procedure currently in place at the Air Force Technical Applications Center (AFTAC). I wish to thank AFTAC for their support in providing this topic and the data. I am especially indebted to Captain Russell Tinsley, my contact at AFTAC, for his aid and support in this effort.

My sincere thanks go out to my advisor, Dr. Kirk A. Mathews for his continued guidance throughout this project and his foresight to allow this project to take place. I am indebted to Dr. Peter J. Rousseeuw, upon whose work and book my methodology is based. I am especially grateful to Dr. Rousseeuw for providing the PROGRESS software that allowed me to complete this undertaking.

Finally, I am especially grateful to my wife, Linda, and children, Alicia and Kristopher, whose patience and understanding guided me throughout this long task.

Keri Robinson

Table of Contents

List of Figures

List of Tables

List of Symbols

g	autocovariance coefficient
H_0	null hypothesis
H_1	alternate hypothesis
n	number of points in the data set
p	number of predictor variables
r_i	residual determined from least squares
R or R^2	correlation coefficient or coefficient of determination
s^o	scale estimate
σ	sample standard deviation
θ_0	constant or intercept of fitted regression
θ_1	slope of fitted regression
w_i	weight determined from least squares
y_i	observed data point
$\boldsymbol{y}$	vector of observed data points
$\bar{y}_i$	smoothed data value (3-point running average)
$\Delta\bar{y}_i$	first difference of $\bar{y}_i$
$\Delta^2\bar{y}_i$	second difference of $\bar{y}_i$

Abstract

This thesis examines the feasibility of using least median of squares (LMS) procedure applied to a reweighted least squares (RLS) autoregression model to identify significant outliers in time series data. The time series were analyzed for data points that were outliers. In order to perform detailed analysis on an outlier, the analyst must be able to determine that an outlier data point is significantly different from normally distributed data. This thesis examines a new method for identifying these outliers.

Data from the field were characterized and fit with time series models using an autoregressive reweighted least squares routine (ARRLS) derived from the LMS methodology. Various orders of autoregression were applied to the ARRLS method to determine an appropriate order for the model; resulting fit coefficients were tested for significance. Regression results from data taken at five sites are presented.

By using an autoregressive order of one (AR(1)) applied to the ARRLS, this method significantly improved outlier detection in the time series data over the recursive removal without regression (RRR) method currently in use. In addition to identifying the outliers found by RRR, the AR(1)-RLS method routinely identified four times as many outliers as AFTAC's RRR method. The AR(1)-RLS method is recommended as a complimentary procedure to the RRR method currently used in identifying significant outliers. After sufficient operational experience is gained, AR(1)-RLS may supplant current schemes. Recommendations for improvements to the AR(1)-RLS method are offered.

IDENTIFICATION OF SIGNIFICANT OUTLIERS IN TIME SERIES DATA

I. *Introduction*

Research Problem

In recent years, the Air Force Technical Applications Center (AFTAC) has sponsored studies to investigate methods to improve its capability to identify significant outliers in time series data. Outlier identification plays a central role in many of AFTAC's efforts. Currently, the analysts use a recursive removal technique developed by AFTAC to identify the outliers. Some data analysts at AFTAC suggest by using this method of outlier identification, certain events that may be significant (but do not meet the strict three-σ criteria) often go undetected. Graphical representation of the time series reinforces this concern. These graphs show apparent outliers in the data that do not meet the criteria that identify them as outliers.

A new method proposed by Dr. Lloyd Currie, of the National Institute of Science and Technology (NIST), tries to solve this problem by using a derivative method that is based on statistical process control theory. He proposes using a three-sample data smoother and identification of outliers by use of z-scores. The calculation of the z-scores is based on the first and second differences of smoothed data.

Unfortunately, the derivative method also performs poorly in identifying some obvious ou s.

These methods, as well as others, fail to identify obvious outliers in the data. A new, robust method for outlier identification is required. The research presented in this paper attempts to solve this identification problem.

Research Objective

The objective of the research is to characterize the data, develop a robust method to detect outliers in the data, and compare the results with other methods currently in use. A robust method is relatively insensitive to the presence of the outliers it is attempting to identify. The aim of this research is to provide the analyst with an additional statistical tool to identify significant outliers in time series data.

Scope, Limitations, and Assumptions

This effort is concerned with time series data. This thesis is limited to the following:

1. Implement an Autoregressive Reweighted Least Squares (AR(1)-RLS) algorithm for the identification of significant outliers in time series data.
2. Benchmark the procedure with actual data sets. Determine the minimum adequate order of autoregression for these data series.

3. Perform a comparison of the derivative algorithm, the AFTAC RRR method currently used in outlier identification, and the AR(1)-RLS method.

The final product will be a test methodology utilizing AR(1)-RLS, which is capable of identifying outliers in time series data.

For this thesis, the following assumptions apply: all data in the time series are discrete data samples, drawn at uniform sampling intervals; massaging of the data to account for missing data points will not be performed, the method of analysis itself will handle a finite number of missing data points; the data do not approximate a normal distribution, but contain long tails of outliers; non-outliers may be non-normal. While the data points will have some measurement errors, uncertainty estimates will not be used in the analysis. Measurement errors are less than one percent and are negligible compared to the time series variations.

Organizational Overview

Chapter Two describes the type of data analyzed. The types of analytical tools available for analysis of time series data are also discussed. Attention is brought to the AFTAC method of data analysis as well as other proposed methods. Finally, the Reweighted Least Squares (RLS) technique is introduced and its merits discussed. Chapter Three begins with Rousseeuw's development of the reweighted least squares procedure by first discussing the problems encountered with using a conventional least mean of squared residuals fit when outliers are present. This is followed with the development of the least median of squared residuals procedure that produces weights that can be applied to the least squares process to produce a robust method for identifying outliers and fitting data with outliers.

Chapter Four develops the test methodology that is the basis for determining the appropriate order of autoregression for the outlier detection model. In this chapter, the development of the graphical methods for displaying the data is presented. Problems with the least squares method and why the development of the reweighted least squares was necessary are expounded. Following the method development sections, the available commercial software that implements the LMS routine is discussed. In Chapter Five, the test methodology of Chapter Four is applied to the example data sets provided by AFTAC and the results are discussed. Confidence tests and goodness of fit tests are performed to determine the appropriate order to be used in the model. AR(1)-RLS, AFTAC RRR method, and derivative method results are compared. The conclusions and recommendations of the thesis follow in Chapter Six.

II. Background

A major part of work being performed at AFTAC involves detecting a significant signal or event out of a noisy background environment. The AFTAC analyst needs a preliminary identification that a significant amount of a radionuclide was released into the environment. This significant amount is called an outlier. It is these outliers that interest the analyst. An outlier is significant if its value is above a background or baseline value. A background level is calculated for each series of data and is therefore series specific. Measurements of the radionuclide in the environment are taken on a daily basis. Throughout this thesis, the recorded value of the radionuclide is referred to as the K-value. The objective here is to determine when a particular concentration of the radionuclide or K-value in the environment is significantly elevated above the calculated background value.

The current outlier identification procedure involves selecting a data population centered on a particular data point. Using this population, an average value is computed. The number of standard deviations that data point is above the average value is determined, and any data point above three-σ is rejected. This rejection is performed until no points exceed three-σ. This final average is called the background value. Finally, detailed analysis is performed on any value in the population that is more than three-σ above the calculated background value. As will be shown, this method is flawed and often fails to identify some obvious outliers.

Introduction

This chapter describes the basics of a time series. The types of time series, the makeup of our data as it relates to a time series, and specific notation used in time series analyses are reviewed.

The methodology currently in place at AFTAC to identify outliers in time series data as well as the problems associated with this approach are discussed. This is followed by a discussion of other proposed methods of analysis to find significant outliers in the time series data.

Finally, the methodology based on the reweighted least squares technique and the advantages it offers in detecting significant outliers are introduced.

Time Series Analysis

Description of the Time Series. In its basic form, a time-series is no more than a set of data $\{y_t:t=1,....,n\}$ in which the subscript t indicates the time at which the data y_t was observed. Diggle categorized time series data as follows:

1. The points in time at which the observations are taken are not equally spaced. The notation for this type of data is $\{y(t_i):i=1,....,n\}$.
2. Each data point represents an accumulation of some quantity over a specified interval of time, rather than its value at a single point. Daily rainfall totals fall into this category.

3. The data set may be augmented by replicate series. Control groups where the same data is taken over a specified period of time fall into this category.

4. Each scalar quantity y_t might be replaced by a vector $\mathbf{y}_t = (y_{1t},\ldots,y_{pt})$ giving the values of *p* quantities which are in some way related. An example of this type might be a daily reading of the temperature, blood pressure, and pulse rate of a hospital patient. (Diggle 1990:1)

The type of time series of interest for this research topic is that of the second category above. Namely, data that are accumulated over the course of a day and reported as a single measurement. Figure 1 is a graphical example of this data. This figure represents nearly two years of data from one site and is typical of the type of data to be analyzed.

An important aspect in time series analysis is stationarity of the data. Most research work in time series analysis has been concerned with the properties of stationary time series. However, if the series is not stationary, then various techniques can be used to remove obvious trends from the series. The most common method to remove trends from a series is differencing. Differencing is used extensively in the derivative method discussed later in this thesis. Jenkins went on to separate time series data into three broad categories based on stationarity:

1. Those which are stationary over relatively long periods of time because of some form of control over external conditions.

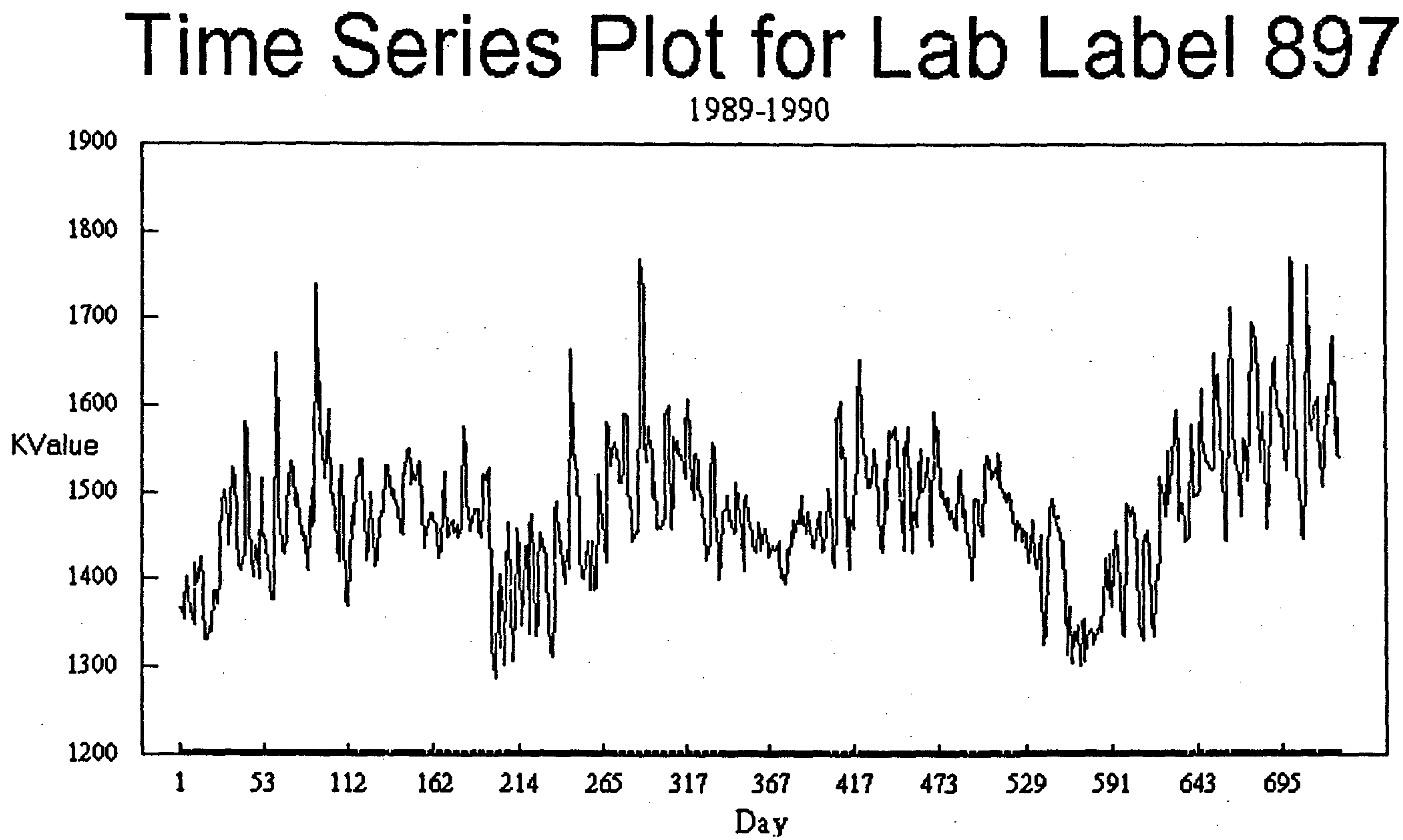

Figure 1. Plot of Data from Site 897.

2. Those series which may be treated as stationary provided a sufficiently short length of series is examined.

3. Series which are quite obviously non-stationary, both from their visual appearance, and also from *a priori* knowledge of the phenomenon being studied (Jenkins, 1968:151).

Examination of the type of time series of concern to AFTAC and to be evaluated in this thesis suggests it is of the second category described by Jenkins. For the majority of the data analysis performed, time periods of only 30 to 90 days are analyzed. This will allow us to consider data with some non-stationarity characteristics to be stationary.

Robustness. Hoaglin *et al* discuss the idea of robustness and the related notion of resistance. Robustness generally implies the insensitivity of a regression procedure to wild outliers (Hoaglin and others 1983: 2) Robustness is a necessary quality in any method designed to identify outliers.

Martin and Yohai discussed the concept of robustness and its importance in performing time series analysis. They stated that a robust procedure should be applied in the time series setting. They regarded qualitative robustness as paramount. By their definition, an estimate is robust when it changes by only a small amount when the sample is changed by replacing a small fraction of observations by arbitrarily large outliers (Martin and Yohai 1985: 120-126).

Breakdown Point. In addition to the idea of robustness described above, another important concept is that of the breakdown point. F.R. Hample introduced the idea of breakdown in 1971. In its basic form, the idea of the breakdown point of a regression estimator is the largest fraction of data that can be moved to infinity without taking the value of the estimate to infinity. The sample mean has a breakdown point of zero, implying that moving a single data point to infinity will drive the mean to infinity. However, the sample median is highly resistant with a breakdown point of approximately ½ for finite sample sizes and tends to exactly ½ as the sample size tends to infinity. The breakdown point is a global measure of performance of an estimator (Martin and Yohai 1985:150-151). It is a quantitative measure of the qualitative property called "robustness".

Hoaglin *et al* further defined the breakdown point of a procedure for fitting a line to n pairs of y-versus-x data as k/n, where k is the greatest number of data points that can be replaced by arbitrary values while always leaving the slope and intercept bounded. A breakdown bound of ½ is the best one can anticipate. Beyond this bound, no distinction can be made between fitting the good data points and fitting outliers (Hoaglin and others, 1983: 159)

Recursive Rejection w/o Regression Method of Analysis

Discussion. The data analysts at AFTAC use a method of analysis provided in an in-house-developed software package called RPP. The AFTAC method is hereafter referred to as the Recursive Removal without Regression (RRR) method. This package provides the analyst with two major methods to view the data, either graphically or in a series of table listings.

RRR Algorithm. The RRR algorithm employed by AFTAC is simple in nature but lacks robustness. The RRR is a recursive routine. The basic algorithm uses a window of data points around a specific day that makes up the sample population. AFTAC uses a window of 30 days based upon statistical minimum population sizes for normally distributed data (Tinsley, 1992).

Simply put, the routine computes filtered statistics (mean, standard deviation, minimum and maximum) on the input data array. The first step is to calculate the number of the non-zero data points in the population. Since AFTAC specifies a population size of 30, the data set consists of the data points 14 days prior to and 15 days after the day of interest. The data points with zero values are first eliminated and the mean of the remaining non-zero values is then calculated.

$$\bar{y}_i = \frac{1}{n}\sum_{j=i-14}^{i+15} y_j \qquad (1)$$

The sum and the sum of the squares of the non-zero points are then calculated.

$$y_{sum} = \sum_{j=n-14}^{n+15} y_j \qquad (2)$$

$$y_{sumsqr} = \sum_{j=n-14}^{n+15} y_j^2 \qquad (3)$$

Once the sum and sum of the squares are calculated for the data set, the standard deviation can then be computed.

$$\sigma = \sqrt{\frac{(y_{sum} - \frac{y_{sumsqr}}{n})}{n-1}} \tag{4}$$

With σ computed, the number of standard deviations each data point in the population is above the mean (or background value) is then calculated. If any point is three or more standard deviations away from the background value, y_{sum} and y_{sumsqr} are decremented by the value and square of that value respectively. Additionally, the number of points remaining, n, is decremented. After all the data points have been screened and those greater than three-σ removed, the σ is recalculated and each of the remaining points is again subjected to the three-σ test. This is repeated until no additional points are removed from the data set or until the number of points remaining in the data set fall below 15.

When the cycle is complete, the mean of the remaining values now represents the background value for that day. This background value is then subtracted from the measured value for that day and the number of standard deviation units is calculated. If the resulting number of standard deviation units is greater than 3.0, the point is considered an outlier and flagged. If more than half the values are missing or excessive (i.e., $n < 15$), no statistics are calculated and no information is available for that data point.

The RRR method is the foundation of significant outlier identification at AFTAC today. However, it is not without its problems.

Most notable is the failure of this method to identify as significant those points in a time series which, when displayed graphically in a time series plot, are obvious outliers. The events at days 91313 and 91328 shown in Figure 2 illustrate this point. This "minor" problem motivates the research being performed here. Figure 2 and Table 1 illustrate the graphical and tabular form of data display produced by the RRR algorithm.

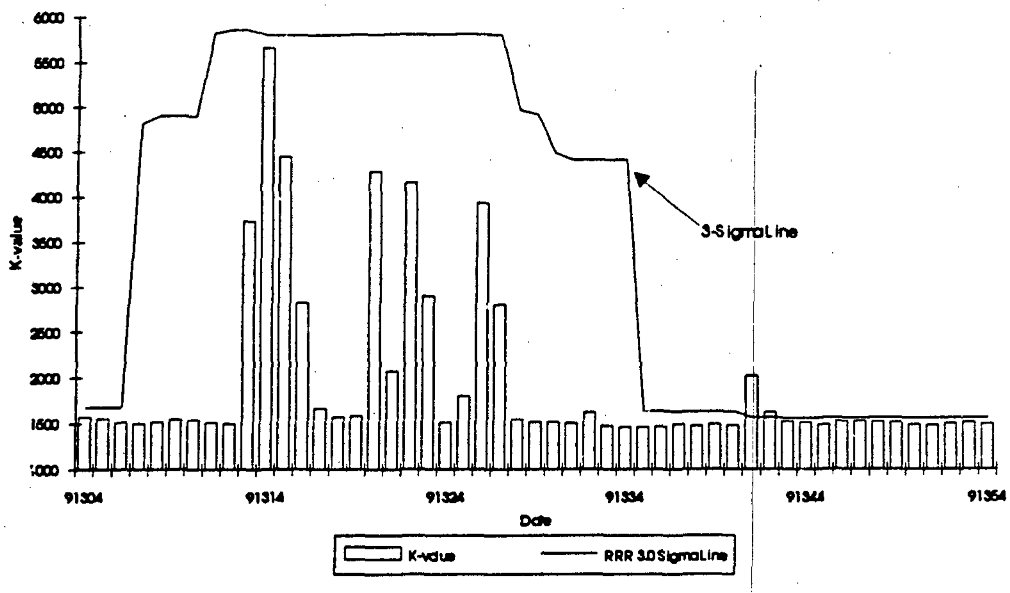

Figure 2. Graphical Display of AFTAC Data

The graphical display shown in Figure 2 gives the analyst a quick look at the site data. Any time the K-value line exceeds the three-σ line, the event is recorded as an outlier. This same information is provided in the listing in Table 1. The listing in Table 1 provides the analyst with a quick look at the station data results. The following information is given in the listing:

DATE	VALUE	BKGND	σ	DRP
91083	6206.2	1949.4	+12.5	5
91084	5090.6	1965.2	+ 9.2	6
91085	9718.9	1965.2	+23.0	6
91086	2608.4	2648.0	0.0	2
91087	1848.5	2740.9	0.6	2
91088	2014.4	2777.2	0.5	2
91089	2423.5	2781.2	0.2	2
91090	2416.1	2730.1	0.2	3
91091	1912.3	2735.4	0.5	3
91092	2328.3	2732.3	0.3	3
91093	1891.9	2730.4	0.5	3
91094	2482.4	2843.4	0.2	3
91095	1961.8	2904.7	0.6	3
91096	1653.4	2891.8	0.8	3
91097	2250.3	2864.8	0.4	3
91098	1860.4	2693.3	0.6	3
91099	6518.5	2564.0	2.9	3
91100	40295.0	2529.3	+28.1	2
91101	6518.4	2505.8	3.0	2
91102	4164.4	2528.0	1.2	2
91103	2570.1	2516.0	0.0	2
91104	1988.5	2486.4	0.4	2

Table 1. Sample Data Listing

DATE Julian date for the observation/sample.

VALUE Actual value in arbitrary units for the observation/sample. (K-value in this thesis)

BKGND The calculated background level based on the surrounding 30 days.

σ The absolute value of the number of standard deviations a value is above the background. Values greater than 3.0 are flagged with a '+'.

DRP The number of data points dropped from the original 30 day calculation.

One major problem with this "quick look" is the ease with which significant data can be overlooked. Additionally, if the analyst must examine large amounts of data, it becomes increasingly easy to overlook an outlier.

Previously Proposed Methods of Analysis

In the past, a number of methods have been proposed to AFTAC in an attempt to better improve the detection of significant outliers in the analysis of time series data, but none have been adopted. For completeness, a brief discussion of four of these methods is included.

The Derivative Method. Dr. Lloyd A. Currie, of the National Institute of Science and Technology, proposed a procedure that demonstrates the derivative method of outlier detection in a background time series. The algorithm is based on five operations, applied to the original data set: 1) interpolation of missing days not to exceed three days, 2) application of a three-day moving average, 3) taking of first differences, 4) taking of second differences (repeating operation-3 on its output), and 5) applying a control process routine to spot out-of-control points (possible outliers), using "local" rather than "global' standard deviation.

Dr. Currie explained that the control limits for this procedure are set to ± 5 standard deviations for sound statistical reasons. In effect, the probability (1-sided) of exceeding these limits by chance, once, in a 365 day set, is approximately 10% for the first differences and 5% for the second differences, when there is random normal error only (null case). Therefore, any excursion you see in the plot of the second differences should be scrutinized as a possible outlier. Significant excursions in the second difference must be negative (negative curvature for a positive outlier), and must be beyond the control limit of -5.0 ("z-score") (Currie 1992).

Dr. Currie provided the following as the pseudo-code for the derivative method for outlier detection. Appendix B contains the algorithm coded in BASIC.

Step-1: Isolate the time series data vector, length-n, to be studied. If it has missing value sequences exceeding length three, break it into sub sequences which do not. Given the time series, y_i, $1 \le i \le n$, the smoothed series $\bar{y}_i$ is

$$\bar{y}_i = \frac{(y_{i-1} + y_i + y_{i+1})}{3}, \quad \text{for } 2 \le i \le n-1. \tag{5}$$

Step-2: Create a first difference vector, by operating on the smoothed series $\bar{y}_i$. The first difference, $\Delta\bar{y}_i$, is

$$\Delta\bar{y}_i = \bar{y}_{i+3} - \bar{y}_i, \quad \text{for } 2 \le i \le n-4. \tag{6}$$

Step-3: Repeat step-2, this time operating on the first difference vector, resulting in a second difference vector. The second difference, $\Delta^2\bar{y}_i$, is

$$\Delta^2\bar{y}_i = \Delta\bar{y}_{i+3} - \Delta\bar{y}_i, \quad \text{for } 2 \le i \le n-7. \tag{7}$$

Step-4: Perform an ordinary control chart operation on the individual elements of first and second difference series, where the "group size" is unity. Compute the mean value for each series as the sum of the elements divided by n-5 or n-8, as appropriate (Dr. Currie used n-3 and n-8 respectively, but this is incorrect). Ideally, the expected values of these means would be zero. The mean of the first and second difference series, $\overline{\Delta\bar{y}_i}$ and $\overline{\Delta^2\bar{y}_i}$, are

$$\overline{\Delta\bar{y}_i} = \frac{\sum_{i=2}^{n-4}\Delta\bar{y}_i}{n-5} \tag{8}$$

$$\overline{\Delta^2\bar{y}_i} = \frac{\sum_{i=2}^{n-7}\Delta^2\bar{y}_i}{n-8}. \tag{9}$$

Next, estimate the "within" or "local" standard deviation ("process-σ"), using the simplest approach, the range technique. Compute the sequence of ranges as the differences between each pair of elements The range of the first differences, $\bar{R}_1$, is

$$\bar{R}_i^1 = \Delta\bar{y}_{i+1} - \Delta\bar{y}_i, \quad \text{for } 2 \le i \le n-4 \tag{10}$$

The range of the second differences, $\bar{R}_2$, is

$$\overline{R}_i^2 = \Delta^2\overline{y}_{i+1} - \Delta^2\overline{y}_i, \quad \text{for } 2 \leq i \leq n-7 \tag{11}$$

Next, compute the average absolute range, as the sum of the absolute values of the differences (ranges) divided by the range vector length.

$$\overline{R}_1 = \frac{\sum_{i=2}^{n-4} \overline{R}_i^1}{n-5}, \tag{12}$$

and

$$\overline{R}_2 = \frac{\sum_{i=2}^{n-7} \overline{R}_i^2}{n-8}. \tag{13}$$

The statistical factor 'd_2' (1.128) converts the mean ranges (for observation pairs) to estimated standard deviations. This gives an estimated σ for each range of differences, σ_1 and σ_2 as

$$\sigma_1 = \frac{\overline{R}_1}{1.128} \tag{14}$$

and

$$\sigma_2 = \frac{\overline{R}_2}{1.128}. \tag{15}$$

This mean range divided by 1.128 gives an estimate of the "process σ." (Ryan, 1989: 84-85, 343).

Step-5: Finally, compute the vector of "z-scores" for the first and second differences. The z-score, z_i, is

$$z_i^1 = \frac{\Delta \bar{y}_i}{\sigma_1}, \quad \text{for } 2 \leq i \leq n-5 \tag{16}$$

and

$$z_i^2 = \frac{\Delta^2 \bar{y}_i}{\sigma_2}, \quad \text{for } 2 \leq i \leq n-8. \tag{17}$$

For reasons discussed above, Dr. Currie sets the control limits for z at ± 5 for the first difference and at - 5 for the second difference. Dr. Currie went on to explain that this procedure is specifically designed to look for outliers that occur as the result of a 'local incursion' and is not valid for 'long range events' which cannot be accurately predicted by the model. (Currie, 1991:1-2).

The STL Procedure. STL is a filtering procedure for decomposing a seasonal time series into seasonal, trend, and remainder components. STL has a simple design that consists of a sequence of applications of the LOESS smoother. The simplicity allows analysis of the properties of the procedure and allows fast computation, even for very long time series and large amounts of seasonal and trend smoothing. Other features include: the specification of amounts of seasonal and trend smoothing which range from very small to very large; robust estimates of the seasonal and trend components that are not distorted by aberrant behavior in the data; specification of the period of the seasonal component as any integer multiple of the time sampling interval greater than one. (ENSCO: 1-2)

The LOESS Procedure. LOESS is a nonparametric regression using multivariate smoothing by moving least squares to fit data. Loess estimates regression surfaces by multivariate smoothing: fitting a locally linear or quadratic function of the independent variables in a moving fashion. This is analogous to how a moving average is computed for a time series. Compared to classical approaches -- fitting global parametric functions -- LOESS substantially increases the domain of surfaces that can be estimated without distortion. Also, a useful feature of LOESS is that analogs of the statistical procedures used in parametric function fitting -- for example, ANOVA and t intervals -- involve statistics whose distributions are well approximated by familiar distributions (ENSCO: 1-2).

The LOWESS Procedure. The LOWESS program contains the routines for the classical LOESS algorithm. It smoothes only as a locally linear function of one independent variable, computes the LOESS curve only at the values of the independent variable in the data set, and computes no statistics. According to ENSCO, you can readily use LOWESS for smoothing scatter plots, since it is simple and fast. Smoothing can be carried out for more than one independent variable, the LOESS surface can be computed at any collection of values in the space of the independent variables, and statistics for confidence intervals and ANOVA can be computed. (ENSCO: 2)

STL, LOWESS, and LOESS were not adapted by AFTAC, mainly due to the complexity in their implementation and the manipulation of the data necessary to get it into a form usable by the procedure. In particular, LOWESS and LOESS required the analyst to make subjective inputs into the model. The derivative method is occasionally being used in limited cases, but has not yet been formally accepted. Again, the limitations of the procedure, the problems associated with missing data, and the requirement for 'local incursion' make its widespread use unlikely (Tinsley, 1992).

Limitations

In each of the previous sections that deal with either methods in use or proposed methods, limitations with these methods have been identified. These limitations range from difficulty of use and implementation (with the LOWESS, LOESS, and STL methods) to manipulation of the data (with the derivative method). The most disquieting problem exists with the RRR method. In many obvious cases of outliers in the data set, the method fails to identify these outliers. The ability of the RRR method to identify probable outliers in a data set appears to hinge not only on the magnitude of the outlier, but also on the size of the population the data is drawn from.

This is illustrated in the following example with is the basis for the RRR algorithm. Consider a data set where the value of all points is zero except for one

$$x_i = 0 \quad \forall i(1 \ldots N) \text{except one}, x_0 . \tag{18}$$

The average of the set is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N} x_0 = \frac{x_0}{N}, \tag{19}$$

and the sum squared is

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 = \frac{1}{N} x_i^2 = \frac{x_0^2}{N}. \tag{20}$$

The sample standard deviation, s, which approximates σ, is given by

$$s(\approx\sigma) = \sqrt{\frac{\sum x_i^2 - N\bar{x}^2}{N-1}}. \tag{21}$$

Finally, it can now be shown that the number of standard deviations a particular point is above the mean is a function of the number of points in the population, N, and not the magnitude of the point. This is given by

$$\begin{aligned} \#\,\text{of}\,\sigma &= \frac{x_0 - \bar{x}}{s} \\ &= \frac{x_0 - {}^{x_0}\!/_{N}}{{}^{x_0}\!/_{\sqrt{N}}} \\ &= \frac{N-1}{\sqrt{N}}. \end{aligned} \tag{22}$$

The results of several population sizes are tabulated in Table 2.

Table 2

Results of Outlier Identification Using RRR Method

N	$\overline{X}$	σ	# of σ	outlier Identified?
5	0.200	0.447	1.789	no
7	0.143	0.378	2.268	no
9	0.111	0.333	2.667	no
15	0.667	0.258	3.615	yes
30	0.033	0.183	5.295	yes

The information given in Table 2 is:

N sample population size,

$\overline{X}$ arithmetic mean of the population,

σ sample standard deviation,

# of σ number of standard deviations above the mean the suspected outlier is,

outlier identification as an outlier.

As Table 2 illustrates, determining whether the data point is identified as an outlier is strictly dependent on the population size. In each case, the data point was an obvious outlier, but the population size was the determining factor in its identification. This is a major flaw in this method.

As a result of the problems discussed for the various methods, it is necessary to develop a methodology for identifying outliers which is not influenced by the population size or the occurrence of the outliers in the data set. To remedy these problems, a robust method with a high breakdown point is required. This method should not depend upon the population size, nor should it be influenced by the presence of the outliers it is attempting to identify. In the next chapter, the AutoRegressive Reweighted Least Squares (AR(1)-RLS) method is developed. The application of the methodology shows great promise in correcting shortcomings in the previously discussed methods.

Summary

This chapter discussed the basics of a time series and how the data that AFTAC analyzes falls into the two general categories of time series--each point represents an accumulation of some quantity over a specified interval of time and that series of sufficiently short length can be treated as stationary. The RRR method currently in place at AFTAC as well as a number of other proposed methods for analyzing the data was discussed. Finally, the need for a more robust method for detecting outliers was identified. In the next chapter, the AR(1)-RLS methodology is developed and followed with the application of the method to detecting outliers in the time series data.

III. *Theoretical Development*

Introduction

In his book *The Analysis of Time Series*, Chatfield writes about graphical displays of time series data:

> The first step in analyzing a time series is to plot the observations against time. This will show up important features such as trend, seasonality, discontinuities and outliers (Chatfield, 1984:14).

Not only is much explanatory information gleaned from the initial look at the graphical display of data, but it also enables the analyst to see the behavior of the data, to see unexpected features as well as the familiar regularities. The emphasis on the visual display of data provides a major contribution to exploratory data analysis (Hoaglin, 1983:3-4).

AFTAC is searching for additional tools to provide the analyst an improved capability to identify significant outliers in time series data. This chapter will begin with the development of techniques and methods for dealing with time series data, including initial identification of outliers by graphical displays. If possible, the analyst would like to examine all data graphically, but AFTAC does not have the resources to do so. Thus, what is needed is a method of reliably flagging outliers so the analyst can later examine the data graphically and decide on further analysis to be performed. The graphical method then is followed by a discussion of robust estimators, and the need for a high breakdown method. Robustness and the breakdown point are important because an efficient method of outlier detection requires that the method itself not be influenced by the presence of the outliers. The drawbacks of using the least squares method are presented with graphical examples. Finally, Rousseeuw's least median of squared residuals method and how it is used to perform the reweighted least squares routine is developed. This chapter ends with a brief discussion of available codes that contain the RLS algorithm.

Procedure Development

Graphical Display. In the book *Understanding Robust and Exploratory Data Analysis*, Hoaglin, Mosteller, and Tukey discuss the four themes of exploratory data analysis. These are resistance, residuals, re-expression, and revelation (Hoaglin and others, 1983:2). It is this revelation through the graphical display of the data that the analyst is looking for and which should be the basis for any further analysis. Much work and computational effort can be saved by the prudent use of various graphical displays of the data to initially identify suspicious trends in the data. Chatfield goes on to say, "Anyone who tries to analyze a time series, without plotting it first, is asking for trouble. Not only will a graph show up trend and seasonal variation, but it also enables one to look for 'wild' observations or *outliers* which do not appear to be consistent with the rest of the data" (Chatfield 1985:7).

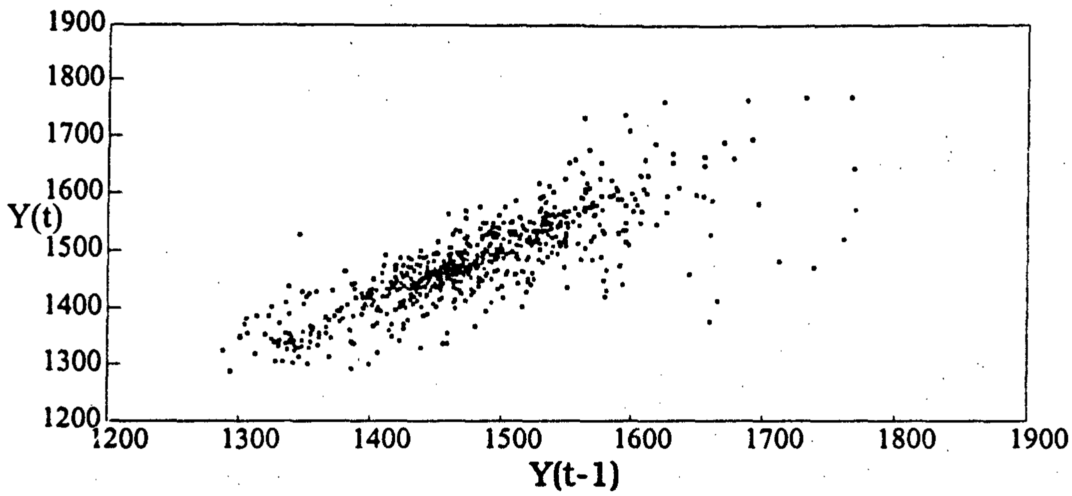

Figure 3. Scatter Plot of Lag of Data vs. Daily Value for Site 897

So important is the graphical display of the data in identifying significant outliers and discovering trends, that two previously discussed methods that are graphically based, LOESS and LOWESS, were suggested as complimentary methods for the analysis of AFTAC's data. By plotting the data, significant trends in the data are discovered. An example would be the scatter plot of the data from one of the sites shown in Figure 3. This is a scatter diagram for lag $k = 1$, obtained by plotting y_t versus y_{t-1}. The plot shows that neighboring values of the time series are correlated, with the correlation between y_t and y_{t-1} being positive. The use of a scatter plot often allows the analyst to better visualize the data structure and identify outliers in either the x or y direction (Rousseeuw, 1987:3). Other plots such as time series plots provide valuable information. Using a time series plot, suspected outliers as well trends in the data can be identified.

Further, index plots, where the standardized residual plotted versus the index of the observation, and residual plots, where standardized residuals are plotted versus the estimated value of the response, are tools for spotting outlying observations. Examples of residual plots are shown in Appendix D in the output from the PROGRESS code. Analysts would use these plots after application of a regression (or autoregression) fit to the data. In addition to the identification of outliers, residual plots can provide a diagnostic tool to gauge the goodness of fit of the model being applied (Rousseeuw, 1987:55-56).

Edward Tufte, in his book *The Visual Display of Quantitative Information*, gives a revealing example of how important it is to graphically display the data. Listed in Table 3 are the data Tufte describes as Anscombe's quartet. All four of the data sets are described by exactly the same linear model, and have identical goodness-of-fit statistics.

Table 3

Anscomb's Quartet (Anscombe, 1973:18)

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

The statistics for these data sets are identical. The mean of the X's is 9.0 and the mean of the Y's is 7.5. The equation of the line for all four sets is $Y = 3 + 0.5X$ and the standard error of the estimate of the slope is 0.118. The total sum of squares $\sum_i (x - \bar{x})^2 = 110.0$, t=4.24, the regression sum of squares = 27.50, the residual sum of squares of Y = 13.75, the correlation coefficient = 0.82, and $R^2 = 0.67$. It is not until you examine a graphical display of the data as given in Figure 4 that it becomes vividly clear how different the data are (Tufte, 1983: 13-14). It is for exactly this reason that the first step in the analysis of any set of data is to graphically display it. Data analysis cannot be performed by simply looking at the statistics alone.

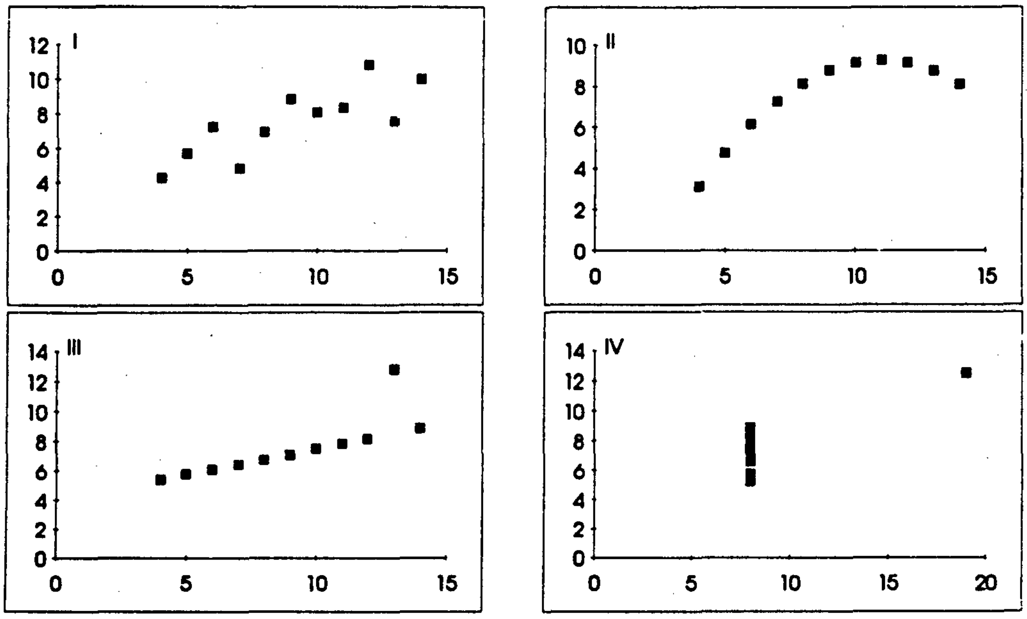

Figure 4. Graphs of Anscomb's Quartet (Anscombe, 1973:19)

Problems with the Least Squares (LS) Regression

Various methods have been developed for fitting a straight line in the form

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i \tag{23}$$

to the data in the form of (x_i, y_i), i=1,....,n. Here θ_0&θ_1 are unknown coefficients to be estimated and ε_i are independent, identically distributed (iid) normally distributed errors. Least Squares (LS) regression operates by minimizing the sum of the squared residuals. It should be noted that minimizing the sum of the squared residuals also minimizes the mean square residual. Thus, LS is really least mean of squared residual regression. This is given as

$$\underset{\theta_0, \theta_1}{\text{minimize}} \sum_{i=1}^{n} r_i^2, \tag{24}$$

where

$$r_i = \hat{y}_i - y_i \tag{25}$$

and

$$\hat{y}_i = \theta_0 + \theta_1 x_i . \tag{26}$$

The reasons for its popularity include ease of calculation, a rather simple mathematical derivation, and that it is built on the Gaussian distribution. Unfortunately, the least squares regression offers no resistance to outliers. In other words, it is not robust. A single wild data point can easily influence the fitted line and cause an erroneous summary of the data. Figure 5 illustrates this point.

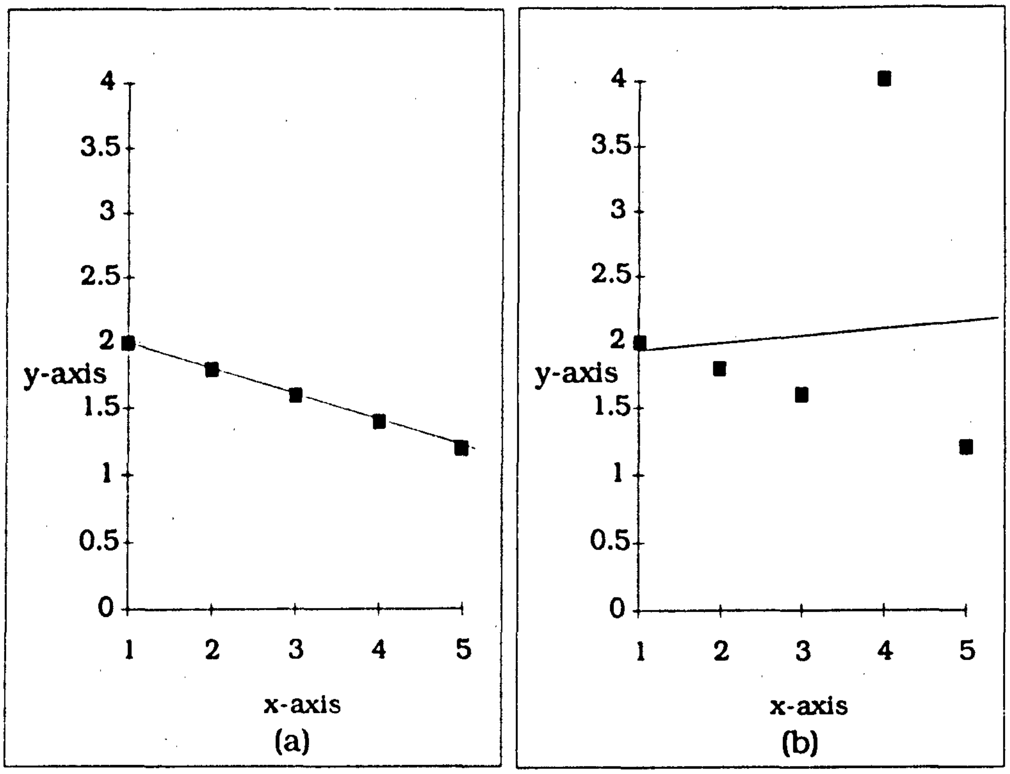

Figure 5. (a) Original data with five points and their least squares regression line. (b) Same data as in part (a), but with one outlier in the y-direction. (Rousseeuw, 1987: 4)

Figure 5(a) illustrates a simple set of data with an LS line fit. If one data point is bad, as in (b), the LS fit no longer represents the data. The LS procedure tries to fit the outlier, even though it is no longer a valid part of the data set. For this reason, a more robust method of fitting the data without being influenced by outliers was necessary.

The Least Median of Squared Residuals (LMS) Algorithm

In the previous section, the classical method of performing a linear regression, the least squares regression, was discussed. Many have tried to improve upon the robustness of the classical regression by replacing the square with some other quantity. One of the first attempts was made by Edgeworth in 1887. It consisted of taking the least absolute value of the residuals and minimizing this sum. This is given as

$$\underset{\theta_0,\theta_1}{\text{minimize}} \sum_{i=1}^{n} |r_i| \tag{27}$$

This technique is often referred to as the L_1 regression, where least squares is the L_2 regression (Rousseeuw and Leroy 1987: 10). While more robust than LS, it was found that the mean was not as robust as the median.

Rousseeuw developed a different approach in which the sum (mean) is replaced by the median of the squared residuals. In light of the median being very robust, this method proved extremely successful. This new robust estimator can handle up to 50% of the data being contaminated (Rousseeuw 1984: 871-872). This least median of squared residuals (LMS) regression, was introduced by Rousseeuw in 1984 and is given by

$$\underset{\theta_0,\theta_1}{\text{minimize}} \text{ median } r_i^2. \tag{28}$$

Rousseeuw said of LMS:

> The computation of the least median of squares regression (LMS) coefficients is not obvious at all. It is probably impossible to write down a straightforward formula for the LMS estimator. In fact, it appears that this computational complexity is inherent to all (known) affine equivariant high-breakdown regression estimators, because they are related to projection pursuit methods (Rousseeuw and Leroy 1987:197).

Rousseeuw gives a brief discussion of the Projection Pursuit (PP) procedures in his book *Robust Regression and Outlier Detection.* Rousseeuw relates this procedure to discovering the structure in a multivariate data set by projecting these data in a lower-dimensional space and to robust regression (Rousseeuw 1987:143-145).

LMS is however, a highly robust method for fitting a linear regression model. For this regression, consider a true model in the form

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i \quad i = 1,\dots,n, \tag{29}$$

or, for multiple variables,

$$y_i = \theta_0 + \sum_{j=1}^{P} \theta_j x_{j_i} + \varepsilon_i \quad i = 1,\dots,n, \tag{30}$$

where there are P explanatory variables, θ's, and the number of degrees of freedom used in fitting.. In the case presented here, there are p independent or predictor variables. For an arbitrary value θ_j, let

$$r_i = y_i - (\theta_0 + \theta_i x_i) \quad i = 1,\dots,n \tag{31}$$

be the residuals, based on the responses y_i and the observed explanatory vectors $\mathbf{x}_i$. In the case of the time series we are examining here, x is a single vector. For the autoregressive model, x_i is the K-Value at time t_i and y_i is the predicted K-Value at time t_{i+1}. The LMS estimate $\hat{\theta}$ minimizes the median of the squared residuals

$$\underset{i=1,\dots,n}{\operatorname{med}}\{r_i^2(\theta)\} = \underset{i=1,\dots,n}{\operatorname{med}}(y_i - x_i\theta_j)^2. \tag{32}$$

In contrast to the LMS method, the normal least squares estimate $\hat{\theta}_{LS}$ minimizes the mean of the squared residuals

$$\operatorname{ave}\{r_i^2(\theta)\} = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\theta). \tag{33}$$

The previous section explained why the least squares estimate lacks robustness. It was shown how a single data point consisting of the response y_i and the corresponding explanatory variable $\mathbf{x}_i$ can cause $\hat{\theta}_{LS}$ to take on any value in p-dimensional space. This is not the case with the LMS method. LMS still provides good statistical performance despite having nearly 50 percent of the data as outliers.

Figure 5 in the previous section illustrated the lack of robustness. Now let's look at the effect of LMS operating on the data. Rousseeuw gives two examples of the magnitude of the problem caused by a single outlier on $\hat{\theta}_{LS}$ and how a robust method such as LMS can correct this and properly fit the line and identify the outlier. Consider two data sets; the first with a single outlier in the y-direction, and the second with a single outlier in the x-direction. These are in Figure 6(a) and 6(b).

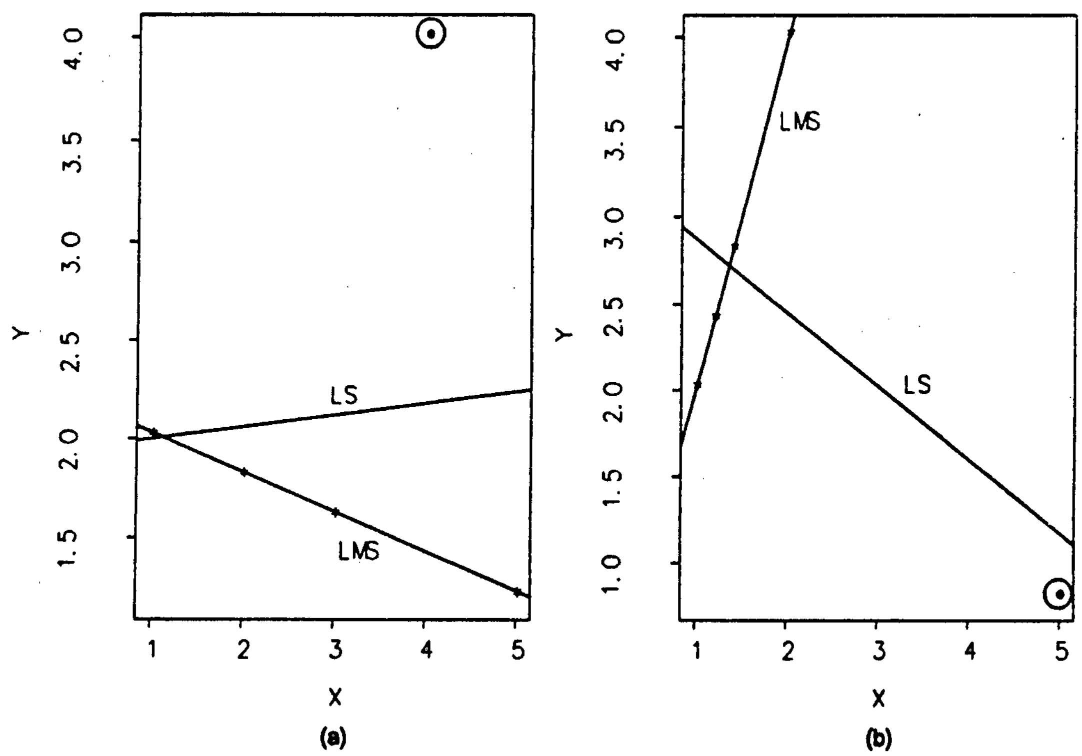

Figure 6. (a) Outlier in the y-direction and (b) Outlier in the x-direction (Rousseeuw, 1987: 4-5).

Figure 6(a) illustrates the best LS fit of a scatter plot of five points that form a somewhat straight line. However, due to a problem, either statistical, copying error, or some other effect, the value of y_4 is incorrect and the point (x_4, y_4) is now far away from the ideal line. As is shown, the LS fit for the data is strongly influenced by this outlier. This point is called an outlier in the y-direction.

Another example is an outlier in the x-direction as shown in Figure 6(b). This outlier is called a leverage point. This is an analogy to the idea of leverage in mechanics. Since x_1 is far from the line, the residual from the LS fit becomes large, and contributes greatly to $\sum_{i=1}^{5} r_i^2$ for the fit to that line. The effect is that the LS line is now tilted toward this leverage point in an effort to reduce this large residual, even though it makes the other four smaller residuals a bit larger. The effect is dramatic (Rousseeuw 1987:5-7). This research will analyze data with outliers in both x an y-directions.

As Figures 6(a) and 6(b) show, LMS is robust, i.e. resistant to these outliers. This is not true with the normal least squares regression, which is strongly affected by the presence of outliers. This is the basis for using a robust regression technique such as LMS to identify outliers in data.

The key feature of the LMS is the robustness that the high breakdown point gives. The breakdown point is approximately 1/2 (and indeed tends to 1/2 as the sample size becomes arbitrarily large). Recall that the breakdown point of a regression estimate is the largest fraction of data that may be replaced by arbitrarily large values without making the estimate tend to infinity.

Weighting and Least Squares

In order to improve the efficiency of the LS method, weighting is introduced. One of the results of the LMS is a scale estimate. The scale estimate is an estimate of the variation of the data, and is similar to the standard deviation. For the LMS, the scale estimate is defined in a robust way. Here it is calculated based on the value of the objective function multiplied by a sample correction factor that is dependent on n and p. Rousseeuw calculates the primary scale estimate using Equation 34.

$$s^0 = 1.4826\left(1+\frac{5}{n-p}\right)\sqrt{\operatorname*{med}_i r_i^2} \tag{34}$$

With this scale estimate, the standardized residuals r_i/s^0 can be computed. The weight can now be calculated for each observation by Equation 35.

$$w_i = \begin{cases} 1 & \text{if} \left| r_i / s^0 \right| \le 2.5 \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

The adjusted scale estimate for the LMS is now calculated using the weights computed in Equation 36. This adjusted scale estimate, associated with Equation 35, is simply the conventional LS scale estimate when the weights are all put to one.

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i - p}} \tag{36}$$

Of particular importance here is that σ^* also possesses the same 50% breakdown point that the LMS method exhibited (Rousseau, 1987: 44,46). This scale estimate is used in the test methodology chapter to help determine the goodness of fit of the coefficients while the weights are used to improve the least squares fit.

Reweighted Least Squares

Using the weights determined by the LMS, a reweighted least squares solution for the data can be found. The effect of using the weights, which can only take on a value of 1 or 0, is the same as deleting all the data points for which w_i equals zero (also referred to as trimmed least squares by some authors). The result would be the ordinary least squares solution if you put w_i equal to one for all cases. The effect of using the weights is to operate on a reduced data set which does not contain outliers. As a result, the statistics are more trustworthy than those associated with the least squares performed on the entire data set (Rousseeuw, 1987: 43-44,132). In the application of this method to the AFTAC data, the remaining data are essentially all background points.

Therefore, the regression, and it's standard deviation, describe only those background points and not the outliers.

To illustrate the improvement in statistical results, Table 4 lists the results from the data used in the PROGRESS run in Appendix D. All of the standard ANOVA results improved, often dramatically, when stepping from the standard LS procedure to the LMS procedure, and finally ending with the AR(1)-RLS procedure. Of importance is the improvement in the σ and R^2 results for the data listed in Table 4.

Table 4

Data Results From PROGRESS Run

	σ	R^2
LS	468.58	0.38
LMS	29.29	0.69
RLS	27.74	0.73

The results in Table 4 demonstrate how much the reweighted least squares, based on the weights determined from the least median of squared residuals, improved the overall statistics. The improvement in σ, which is a measure of the variability of fitted values around the mean is dramatic. Additionally, the R^2 values improved significantly. The higher the R^2, the better the data fit the regression equation.

Available Codes

To test the premise that the LMS and RLS methods could satisfactorily operate on the data sets provided by AFTAC, a version of the LMS method was implemented using Mathematica. With this code, it was confirmed that LMS and RLS could identify outliers in the data. The problem with the Mathematica version was that it was extremely slow due to the overhead of the powerful but interpreted language, as well as the computational complexity of the method. For this reason, a search for commercially available software that incorporated the LMS or RLS method was conducted.

Rousseeuw stated in his preface that the code had been integrated into the workstation package S-PLUS from Statistical Sciences, Inc. I contacted Statistical Sciences and was able to obtain a demo copy of their recent S-PLUS for DOS release. I was able to perform calculations with this product, but found it too cumbersome, mainly because it does not function in the Microsoft Windows environment.

Rousseeuw also stated in his preface that his Program for RObust reGRESSion (PROGRESS) could be obtained directly from him. After exchanging correspondence with Dr. Rousseeuw, he provided a copy of the PROGRESS code (Rousseeuw, 1992). Using the PROGRESS code, the methods developed in the next chapter are tested.

Summary

This chapter looked at the development of the method to be used in detecting outliers. As discussed by many authors, the first step in analyzing any set of data is to display the data graphically.

Development of the RLS process was discussed. By looking at the weaknesses of the original least squares method, and developing the least median of squared residuals method, significant improvement in the detection of outliers was demonstrated. Also discussed was the robustness of the LMS method with respect to outliers in the data set. Additionally, the capability of the high breakdown point to improve the method's capability to withstand up to 50% of the data being contaminated was discussed. Finally, the RLS method was introduced. This robust, high breakdown method was identified as the method of choice for model development.

The chapter ended with a discussion of available codes that incorporate the LMS/RLS methodology for production use. In the next chapter, the methodology to develop a procedure for identifying outliers in a time series is discussed.

IV. *Test Methodology*

Introduction

In this chapter, the test methodology used to determine the most appropriate order of autoregression to use with the reweighted least median squares procedure is developed. Actual data from the last 165 days of 1991 from site 889 is used in this development. I choose this data set because of obvious significant events that are present in the time series plot. In the next chapter, this methodology will then be applied to all data sets.

The tests and methods used in this section are based on developing models for forecasting. Many of the tests for determining order are derived directly from those used to develop Autoregressive Integrated Moving Average (ARIMA) models as described by Box and Jenkins (Box and Jenkins, 1976:18). The test methodology presented here departs from the application of the Box and Jenkins results used in normal forecasting. This method is not trying to predict what the K-value will be on any particular day, but whether that K-value is statistically different from other days around it.

The first step is graphically displaying the data using some common methods employed in time series analysis. Initial characterizations about the data are inferred from the graphical displays. Following this, correlograms -- the autocorrelation and partial autocorrelation functions -- for the data are calculated and plotted. This will give an initial indication of the order of autoregression (AR) appropriate to the data. These orders of autoregression are applied to the data and used as input to the RLS procedure in PROGRESS.

In addition to the PROGRESS runs on the data, stepwise multiple autoregression will be performed. The results from the PROGRESS runs and the stepwise multiple regression will then be used to select the appropriate AR order. This final choice of AR order will then be applied to the data set and outlier statistics calculated.

Test Data

The previous chapter demonstrated the benefits of using a reweighted least median of squares method for fitting a line to the data. This same method can be used for detecting outliers in time series data. The final test of effectiveness of a method is a measure of its performance with actual data. In particular, any new method must be capable of performing as a complimentary process or functioning as a replacement procedure to the existing method.

The data to be analyzed here consist of two years of raw data from six geographically different sites. This data represents sets which range from a stable background with little fluctuation in the data, to extremely noisy data with a large fluctuation in the background. Figure 7 is a time series plot of the data from three sites that are stable, moderately noisy, and extremely noisy.

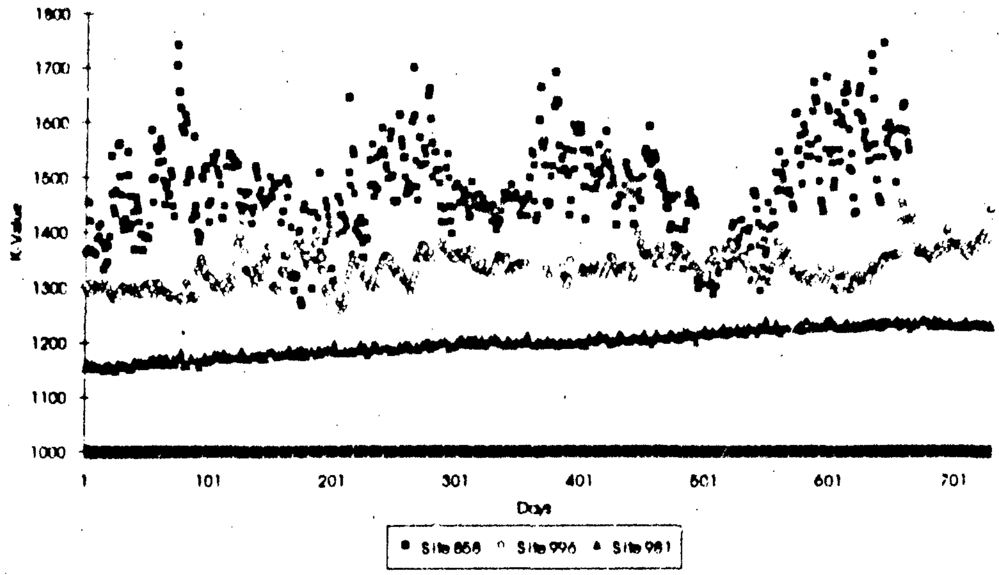

Figure 7. Time Series Plot for Sites 858, 981, and 996

Figure 7 readily illustrates the wide variety of data that is collected and must be analyzed. An effective model developed should be able to span the range of data types illustrated here.

The data used in the development of the methodology were taken from the last 165 days of 1991 for site 889. This data set is listed in Appendix E. This data set was chosen because of what appears to be a smooth time series data stream with possible outliers in the data. These outliers appear near the end of the period of interest. The first step is to graph the data to see whether any significant deviations appear.

Time Series Plots. The time series plot of the data is given in Figure 8. Missing data are indicated on the plot. This plot shows a time series that is relatively flat and stationary with the exception of a few data points that appear to be significant outliers during the period 91313 to

91327. Because the magnitude of the data does not increase or decrease relative to time, a regression technique using time as the explanatory variable and K-Value as the response is not a useful choice. This assertion will be justified later in the analysis.

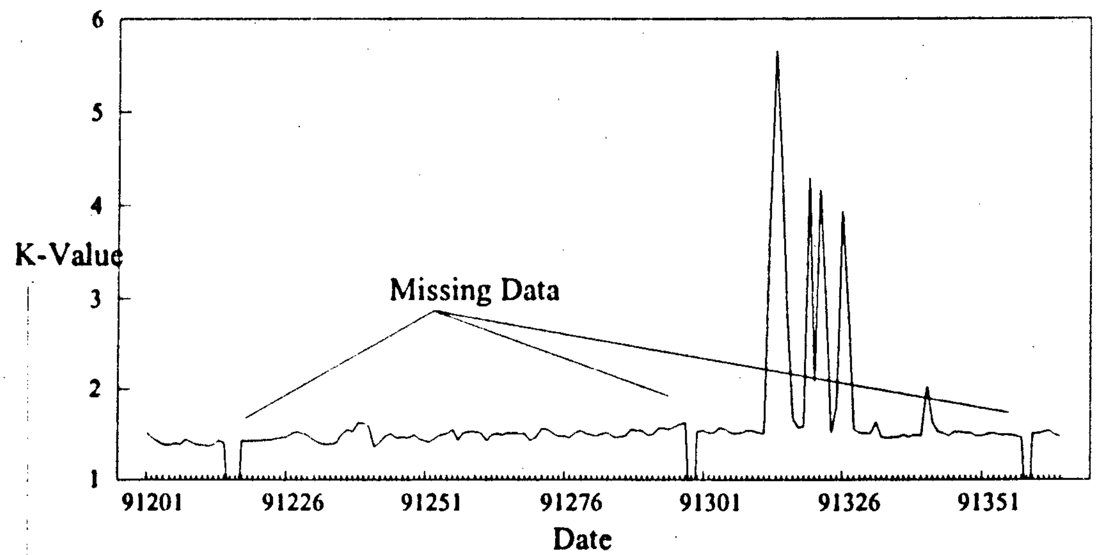

Figure 8. Time Series Plot for Site 889 from 91201-91365

Scatter Plots. Since the model is expected to be a autoregression model, a second way of observing the data is in a scatter plot. This is simply a plot of the of the specified lag value versus the value of a particular day (y_t, y_{t-1}). In a simple regression model, it is easy to visualize the data structure using a scatter plot. In a general multiple regression model with large number of explanatory variables, this would not be possible. Figure 9 is a scatter plot of the data.

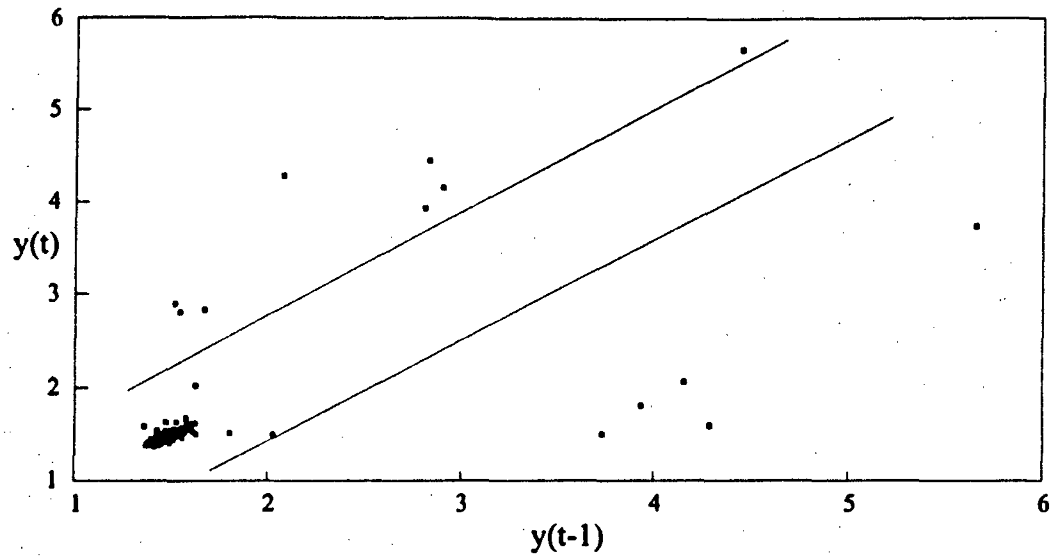

Figure 9. Scatter plot of Site 889 Data.

The general appearance of the scatter plot shows a tight group of points with a slope of approximately one. The two lines drawn give a rough estimation of the trend of the data. Points located above or below the lines should be flagged as probable outliers. Again, since the points appear to fall on a somewhat straight line, the scatter plot indicates that an AR(1) regression model is appropriate. Since time cannot be used as an explanatory variable, the only other choice is some order of autoregression.

The origin of the term autoregressive is taken from the fact that the equation we use to describe an autoregressive model is exactly like a normal regression equation. The difference is, where x_t plays the role of the explanatory variable and y_t the response variable in a regression model, now y_{t-1}, y_{t-2}, etc. are the explanatory variables. Since the variables y_{t-1}, y_{t-2}, etc. are the same data as y_t (just offset by one period, two periods, etc.), y_t is actually being regressed on itself---hence the term autoregressive (Hoff, 1983:50)

Autoregressive Order Identification

The identification stage in determining the order of regression is the longest and most difficult. Fortunately, computers can rapidly produce results based on the methods chosen, but often the identification requires subjective judgment. Once the order of regression is identified, there is relative certainty that the model will be able to accurately fit the data. If the model can fit the data set, it can identify outliers in the data set.

Identification means using the data and any information on how the series was generated to pick a process to begin model generation (Box and Jenkins, 1976:171). A typical key to identification of an AR process lies within the patterns found in the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) (McCleary, 1980:93). The plots of the ACF and PACF functions are commonly referred to as correlograms. Additionally, ANOVA statistics, along with F-tests and t-tests of the coefficients, sample standard deviations, and coefficients of determination (R^2), will aid considerably in the choice of the proper order of autoregression. In particular, the overall F-test will be used to determine whether or not all of the independent variables taken together significantly contribute to the prediction of the dependent variable. The t-test is used to assess whether or not the addition of any specific independent variable to the model significantly improves the prediction of y, given that other variables already exist in the model.

ACF and PACF Plots. Graphical methods, such as the ACF and PACF plots are very useful in the identification stage (Box and Jenkins, 1976:173). Nonstationarity can be recognized by examining either the time series plot, or more commonly, by the graph of the ACF.

For an equally spaced time series $\{y_t{:}t=1,\ldots,n\}$ we use $\bar{y}$ to represent the sample mean. where $\bar{y}=(\sum y_i)/n$, and we define the kth sample autocovariance coefficient,

$$g_k=\sum_{t=k+1}^{n}(y_t-\bar{y})(y_{t-k}-\bar{y})/n. \tag{37}$$

Then the kth sample autocorrelation coefficient is

$$r_k=\frac{g_k}{g_0}. \tag{38}$$

The plot of r_k against k is called the correlogram of the data. Correlograms are often used to check for evidence of any serial dependence in an observed time series. Values of r_k greater than $2/\sqrt{n}$ in absolute value can be regarded as significant at about the 95% level. More often, the correlograms are used to suggest the order for autoregressive models. The reliability of the correlogram for this purpose increases with the length of the time-series on which it is based. (Diggle, 1990: 39-47).

For the majority of the AFTAC data, sufficiently short periods of data are used in the analysis that the data can be considered stationary even if some level of nonstationarity exists (Jenkins, 1968: 151). Hoff writes extensively on using the ACF and PACF plots to identify the order of autoregression. In the book *A Practical Guide to BOX-JENKINS Forecasting,* Hoff gives many examples of the various types of time series one may encounter and the order of autoregression normally applied to that specific data series (Hoff, 1983:54-86). These examples guided the determination of the proper order of autoregression for the AFTAC data sets, although the patterns in the actual data are not as obvious as those in the examples given in the literature. The expected patterns are for infinitely long realizations (McCleary, 1980:94). All the authors suggest that a relatively long series of data is required for time series analysis. Box and Jenkins say at least 50 observations, and preferably over 100 observations, should be used (Box and Jenkins, 1976:18).

The autocorrelation function plot and partial autocorrelation function plot for the Site 889 data set currently under discussion are shown in Figures 10 and 11. The ACF and PACF plots should be viewed together and a judgment made from both (Mykytka, 1991). In the case

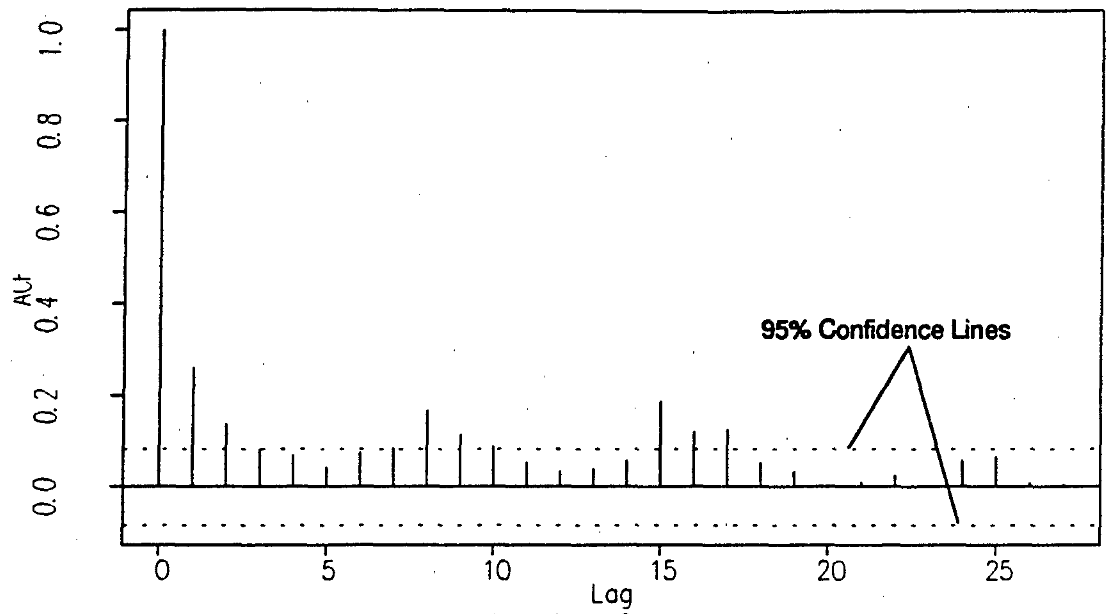

Figure 10. ACF Plot of Site 889 Data

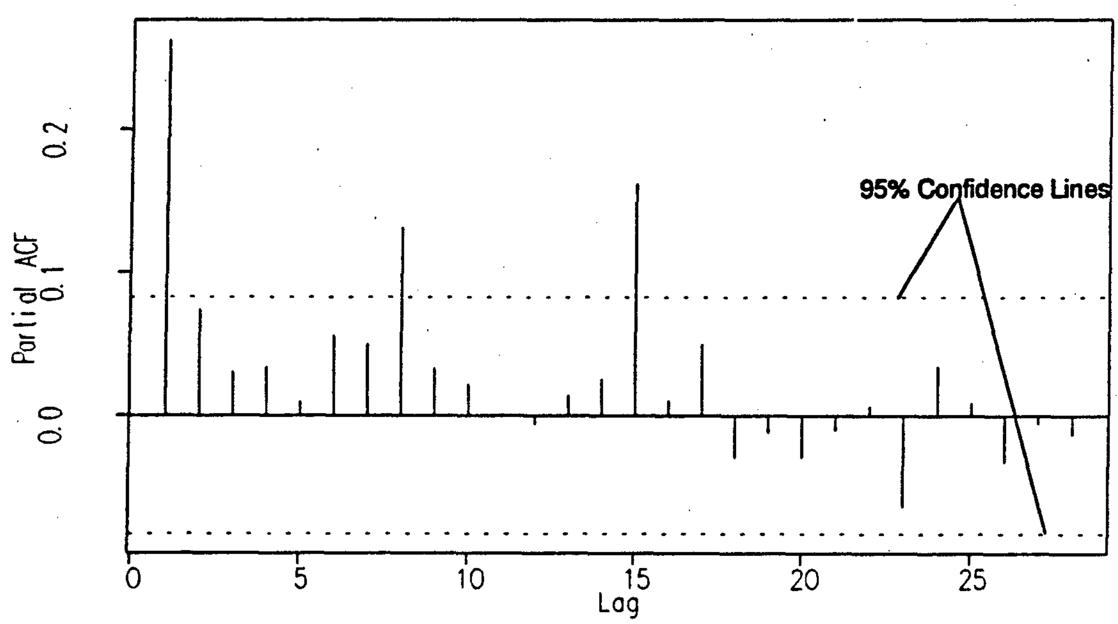

Figure 11. PACF Plot of Site 889 Data

under study, the ACF indicates a strong correlation of the data out to lag two. In this context, strong is defined as significant above the 2.5 σ line. However, when the PACF is used in conjunction with the ACF, it indicates that the correlation is actually only good at lag 1. Other lags show levels of significance above the 2.5 σ line, but are sufficiently far out in lags to reduce their importance to the model. The lag value at lag 8 also shows significance and bears further investigation.

It is necessary to point out that the ACF and PACF plots are just one of many tools used to determine the best order of autoregression for the model. This information will be combined with other results for final formulation. This does however, give an excellent starting place in identifying the order of the model.

Because the two plots are not definitive, two additional techniques to aid in the determination of the AR order are applied. In the next case, the results from the RLS output of PROGRESS are used to provide statistics on which AR order to use.

Confidence Tests. The ACF and PACF results have now given a starting point for final determination of the AR order appropriate for the outlier detection model. The ACF and PACF plots suggest an appropriate AR order. This is then used to test the hypothesis that the coefficients are significantly different from zero.

To determine a confidence level in the regression coefficients, a test based upon the ACF and PACF plots was used. Confidence intervals based upon a Student distribution with $n - p$ degrees of freedom are then applied. For this test a 95% confidence interval is used. The hypotheses are

$$
\begin{aligned}
H_0&:\ \theta_i = 0 \quad \text{(null hypothesis)}\\
H_1&:\ \theta_i \neq 0 \quad \text{(alternative hypothesis)}
\end{aligned}
\tag{39}
$$

This type of test can be helpful in determining if the ith coefficient might be deleted from the model. If the null hypothesis in Equation 39 is accepted for a certain confidence level, the ith coefficient contributes little to the explanation of the response variable and can be removed from the model (Rousseeuw, 1987: 40-41).

The PROGRESS code was chosen as the diagnostic tool to test the hypothesis on the suggested coefficients. Based upon the ACF, a regression model based upon the first, second, and eighth lags may be appropriate. The PACF suggested that only the first and eighth are actually significant in predicting the response variable. Using this information, PROGRESS was run on the data set using a combination of the lags as predictors.

Two statistics which can be used to test the validity of the model are the F and t-tests. For the F-test, the hypotheses being tested is whether the entire vector of regression coefficients, excluding the constant term, equals the zero vector. This is the same as

$$
\begin{aligned}
H_0&:\ \text{All nonintercept } \theta_j\text{'s are together equal to zero}\\
H_1&:\ H_0 \text{ is not true}
\end{aligned}
\tag{40}
$$

The t-test then determines which of the coefficients are necessary. P-values are also computed by the PROGRESS code. The P-value indicates the level of statistical significance of the hypothesis that the predictor variable has an effect on the response variable. It is the

Table 5

P-Values for Regression Coefficients

Variable	Lag 1	Lags 1 & 2	Lags 1 & 8	Lags 1, 2, & 8
Lag 1	0.00000	0.00000	0.00000	0.00000
Lag 2	-	0.00058	-	0.00152
Lag 8	-	-	0.64052	0.59441

probability that the observed fit would occur as a result of random noise in the data. Thus, a small P-value indicates that the fit is statistically significant. An example of the calculation results is given in Appendix D on page 100. The results are presented in Table 5.

The coefficient for Lag 1 was kept in all combinations since both the ACF and PACF plots indicated it was significant. As Table 4 shows, based on the P-values, the coefficient for Lag 8 is not significant at the 95% confidence level and should be eliminated from the model. In both cases where Lag 2 was used, the P-value indicated it was significant at the 95% confidence level.

If the calculated P-value of the associated F distribution is less than the 95% confidence level, then H_0 above can be accepted. If not, it must be rejected (Rousseeuw, 1987:43). Unfortunately, for the cases considered here, the P values associated with the F-test values were all near zero and could be considered valid. Therefore, this test provided no additional information for this data set.

For this reason, additional confidence tests should be performed. Another test that considers how each coefficient individually effects the regression when combined with others is the stepwise regression. Stepwise regression provided the final check of the model parameters.

Stepwise Multiple Regression. In addition to the specific values given above for F-tests, stepwise multiple regression can be used to determine which explanatory variables are significant. Stepwise multiple regression returns only those variables with significant values for the F-test at specified levels.

For this portion, MINITAB software performed the stepwise multiple regression on the RLS output data. The results for Site 889 are given in Table 6, which lists the constant term, the coefficient for each regression term, the T-ratio for each coefficient, and the sample standard deviation and R^2 value for each step. At each step, MINITAB calculates an F-value for each of the explanatory variables given. In the cases evaluated, the explanatory variables were the lag values for autoregressive order one, two, and eight as predicted by the ACF and PACF. If the t-test value of any explanatory variable is less than the specified value of significance, the variable with the smallest F-test value is removed from the model. MINITAB then calculates a new regression, prints the results, and proceeds to the next step. Once the stepwise regression reaches the point where no explanatory variables can be added or removed from the equation, the procedure ends (Schaefer and Farber, 1991: 261-268).

Table 6 lists the constant, T-value, and regression coefficient for each set of variable used. The first step of the stepwise regression calculated the regression with all three explanatory variables in the equation. For the second step, the T-value for the Lag 2 component was too small and it was removed. Finally in the third step, the T-value for the Lag 8 component was below the level of significance and was removed, leaving only the AR(1) component.

Table 6

Stepwise Regression of Site 889

(MINITAB Output)

STEP	1	2	3
CONSTANT	600.9	581.1	625.1
lag1	0.627	0.600	0.618
T-RATIO	7.92	8 98	9.81
lag2	-0.048		
T-RATIO	-0.64		
lag8	0.055	0.047	
T-RATIO	0.92	0.80	
S	470	469	469
R-SQ	38.58	38.42	38.16

While the F and t-test results were inconclusive, the results from the stepwise regression indicated that only the AR(1) component of the data was statistically significant in fitting the data. Based upon these results and those of the ACF and PACF plots, an autoregression of order one applied to the reweighted least squares process proved to be an adequate model of the data examined.

Diagnostic Checking. Having identified the process, the model can now be tested. For this portion, a final run of PROGRESS used the observed data as the response variable and a lag of one day for the explanatory variable. As a final check to the validity of the AR(1) model applied to the RLS procedure, the ACF of the residuals was calculated and plotted. A good model will leave only white noise and has no remaining pattern in the residuals. The ACF will all be insignificant (Makridakis, 1983:446). However, at the 0.05 significance level, a chance does exist for a few significant spikes in the ACF at distant lags (McCleary, 1980:99). The ACF plot for the residuals based on the AR(1)-RLS results from PROGRESS is shown in Figure 12.

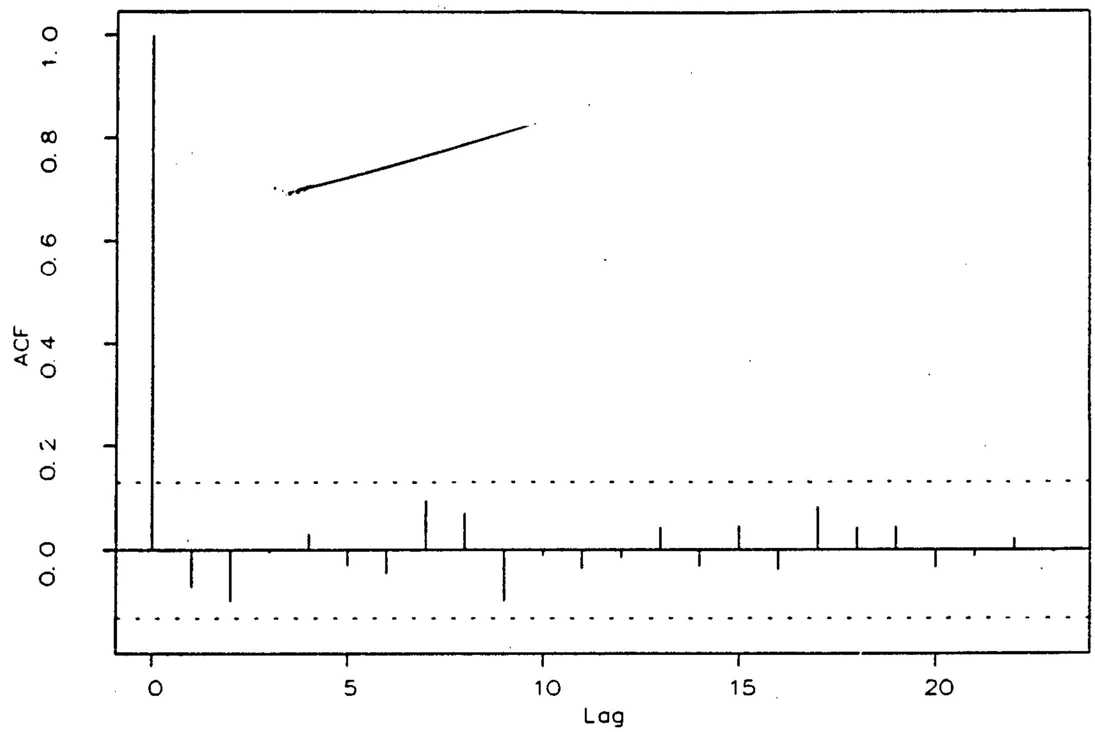

Figure 12. ACF Plot of Residuals from Site 889.

The ACF plot shown in Figure 12 indicates that the residuals are essentially white with no significant spikes. This plot provides strong evidence that the coefficients chosen for the model are significant.

Conclusions

Using the methodology discussed in this chapter, the tests performed indicate that an autoregressive order of one applied to the reweighted least squares procedure provides an adequate model of the data set examined. This data set was fit to insure that the explanatory variables were significant. Coefficients that proved significant were kept while any which were insignificant were dropped.

Goodness of fit tests on all coefficients determined their statistical importance. Additionally, PROGRESS and MINITAB codes calculated ANOVA type statistics for all combinations of coefficients considered in the model. As a last validity check, stepwise regression was performed on the model. Finally, results of residual tests were examined to ensure the residuals left only white noise.

On the basis of the tests performed in this chapter, I concluded that an AR(1) method applied to reweighted least squares was an appropriate model of the data. The next chapter discusses the results of the AR(1)-RLS methodology as applied to five other data sets which AFTAC provided.

V. Results

The results of the data analysis using the methodology developed in the previous chapter are discussed in this chapter. Results are reported for each of the five sites for which data was provided.

The first conclusions are drawn from graphical displays of the site data, including time series plots, correlograms and scatter plots. These will be used to assist in identifying the order of autoregression appropriate for the model. In addition to the graphical displays, confidence test results will be analyzed for the various coefficients of regression. Results from the ANOVA statistics confirm that AR(1) is the appropriate order of autoregression for the final outlier detection model.

Following the discussion of model choice, the outlier results from the AR model are discussed. These results will then be compared with those of the RRR and derivative methods. The main point is not to develop a model that will best fit data sets from one site, but to develop a model that can adequately fit data from any site. How well the AR(1)-RLS model performs in comparison to the RRR and derivative methods will determine its usefulness to the analyst.

For each of the five sites for which AFTAC provided data, the first 300 days in 1989 are analyzed. The analysis was restricted to 300 day blocks by the input array size limitations of the PROGRESS code provided by Dr Rousseeuw. In addition to the 300 day analysis, analysis was performed on the subsets of the data from Site 996 to determine the relative effectiveness of the method when employed on various sizes of data sets.

Analysis of Graphical Displays

The first step in analyzing the data is to display the site data as time series plots. The time series plots for sites 858, 981, and 996 were shown previously in Figure 3. Correlograms, ACF and PACF plots, were created for each site and are shown in Appendix E. These plots were used to provide a first approximation of the order of autoregression to apply to each data set. For each site, the correlograms suggested that an order no greater than three would provide the basis for further investigation into the final order of regression for the model. This decision was made because only the PACF plot from Site 858 had a significant r_k beyond lag three. Again, the idea is to try to fit a model that supports identification of outliers from any site, not just one specific site.

Confidence Tests

As in the previous chapter, the RLS regression was used to provide ANOVA type results on the regression coefficients used in the model. RLS regression was performed on all five sites using lags one, two, and three as the explanatory variables. In addition to the lags, the date of occurrence (t in the time series) was used as the explanatory variable. As expected, time is not a good predictor of future values or in identifying outliers. These RLS results using time as the predictor are presented in Table 7.

Table 7

Regression Results Using Date as Explanatory Variable

For Site 996

Variable	Coefficient	Std. Err	t-Value	P-Value
Julian Date	0.063	0.05734	1.10158	0.27263
Constant	-4301.847	5113.58900	-0.84126	0.40171

Table 7 illustrates that the Julian date is a poor predictor of the K-value for that date. The P-value indicates it is not significant at the 95% confidence level. Additionally, the t-value is not significantly different from zero and the coefficient is extremely close to zero. These results, along with an R^2 value of 0.00904 and a P-value based on the F-test of 0.27 clearly illustrate that the Julian date should not be used as a predictor for the K-value in this model.

The next step in the test methodology was to perform the AR(1)-RLS regression for up to lag three for all five data sets. From this analysis, PROGRESS calculated the R^2, σ, and P-value for the F-test for each order of regression as well as the P-values for each of the coefficients in the regression. These results, along with the stepwise regression to be performed later, provided the best estimate of the autoregressive order to use in the model.

F-Test Results. The group of P-values for the F-test of regression coefficients obtained for each site provided no conclusive results. This is the same problem encountered in the previous chapter. The P-values for all combinations of coefficients for the five sites were zero. This indicated that overall, lag one, lags one and two, or lags one, two, and three tested equally well in predicting the response variable. This meant that the R^2, σ, and the P-values for the individual coefficients would have to be used to determine the order of regression.

Adjusted R^2 Results. The adjusted R^2 results (hereafter referred to only as R^2) from the AR(1)-RLS runs are shown in Figure 13. R^2, or the coefficient of determination, is a measure of the strength of the linear relationship between the response variable and the explanatory variables. R^2 measures the proportion of total variability explained by the regression. In the simple case with a constant term, the coefficient of determination equals the square of the Pearson correlation coefficient (Rousseeuw, 1987: 42). Unfortunately, the results based upon the

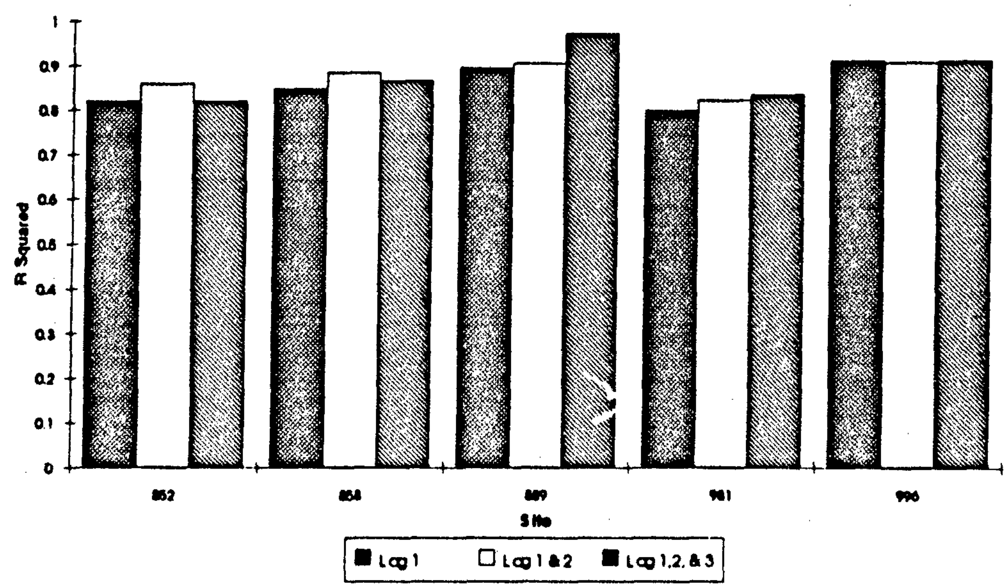

Figure 13. R^2 Results from AR(1)-RLS

adjusted R^2 were inconclusive and other factors need to be examined.

Scale Factors. The next test for goodness of fit is the scale factor (σ^*). The scale factor is a robust version of the sample standard deviation. The best model should provide the lowest scale factor for a given site. Since the objective is to provide a model that performs best overall, we want to minimize the scale factor over all the sites. For each of the sites, the scale factor was calculated for each of the three lag combination regression models. The results are shown in Figure 14.

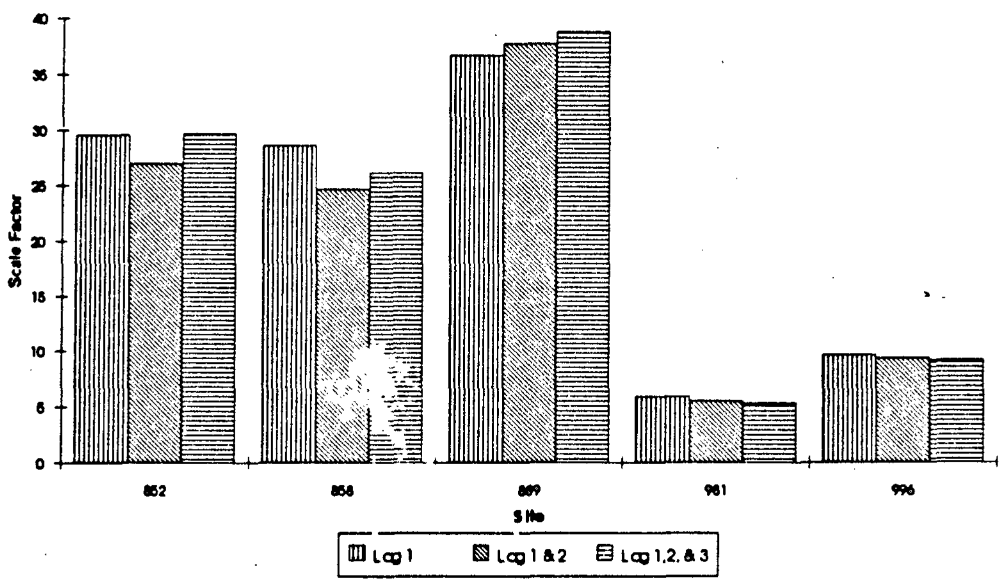

Figure 14. Scale Factors by Site

The results of the scale factors are as unenlightening as those of the R^2 test. Based on these results, no definitive conclusion can be drawn between using any of the three lag combinations. Depending on the site, the difference in scale factors ranged from 2-13%. Therefore, the scale factors indicate that either set of coefficients applied to the AR(1)-RLS model appears to perform about equally.

P-Values of Coefficients. The next test for goodness of fit is to look at the P-values of the individual coefficients. The P-values for the F-test, which fit all of the regression coefficients together, was significant at the 95% confidence level, those for the individual coefficients indicate they should be rejected. For sites 858 and 996, the P-values for the coefficients for lag 2 and for lag 3 all exceeded the 5% level. This implies these coefficients would have to be rejected at the 95% confidence level.

Stepwise Regression. The final test in determination of the coefficients to be used in the outlier detection model is to perform a stepwise regression on the coefficients. This procedure was discussed in the previous chapter. Again, the MINITAB software performed the stepwise regression. For each site tested, lag 2 and lag 3 were systematically eliminated from the regression. Each case left the lag 1 coefficient as the only significant coefficient in the regression model.

Conclusions of Confidence Tests. The final conclusion reached was an autoregressive order one reweighted least squares model (AR(1)-RLS) was the most appropriate model overall. While systematically adding lag 2 and lag 3 parameters to the AR(1)-RLS model gave better results at specific sites, the lag 1 AR(1)-RLS model provided the best overall results that spanned the sites. Furthermore, while the R^2 test, the P-value, and F-test proved inconclusive individually at each of the sites, the stepwise regression clearly indicated that lag 1 was the best choice for the model independent of the site.

Method Comparison

Once the order of autoregression for the AR(1)-RLS model was determined, the test methodology was validated by comparing the AR(1)-RLS results with those of the RRR and derivative methods. The three models, AR(1)-RLS, RRR, and derivative, were run with 300 days of data from each of the five sites. From these model runs, the number of outliers found by each method was tabulated. The results are shown in Figure 15.

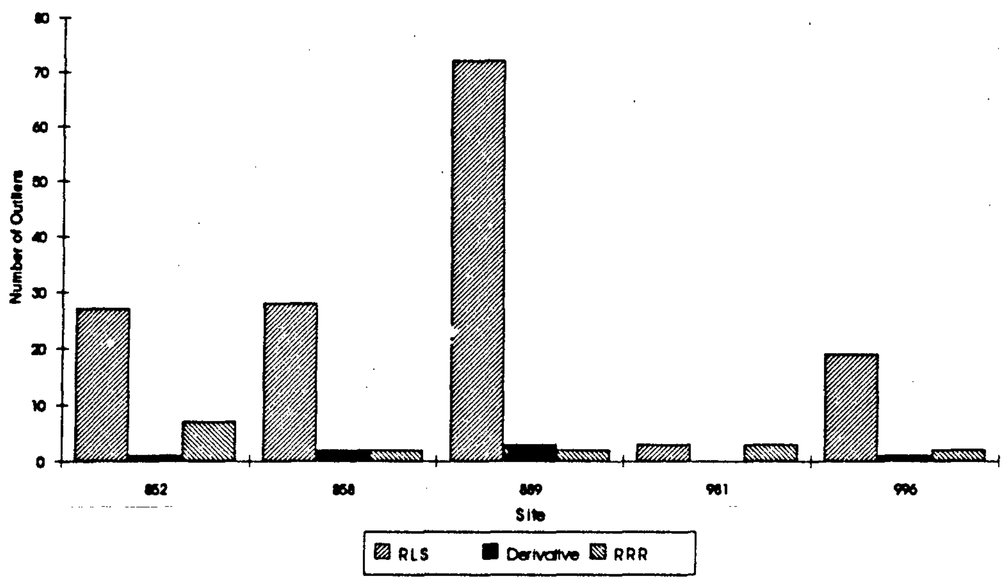

Figure 15. Number of Outliers by Method

For both the RRR and AR(1)-RLS methods, the cutoff for detecting an outlier was set at 2.5 σ. This was done to ensure both models were working at the same level of significance. The normal cutoff for the RRR method is 3.0 σ. The cutoff for the derivative method was $Z_1>5$ (for the first derivative variable) and the corresponding $Z_2<-5$ (for the second derivative variable). This equates to a 95% probability that the event was significant.

In all cases, the AR(1)-RLS method detected more outliers than the derivative or RRR methods. In some cases, such as site 889, the difference was dramatic. For all the site data analyzed, the AR(1)-RLS method found all of the outliers identified by the RRR method.

The time series discussed previously in the test methodology section was again analyzed with the AR(1)-RLS and the RRR models. The AR(1)-RLS model used a cutoff of 2.5 σ and the RRR model used both 2.5 and 3.0 σ. The use of the two different σ values for the RRR method is to show that the method is not particularly sensitive to the two different σ values. Figure 2 on page 13 illustrates the RRR method used at the 3.0 σ level. The next two figures illustrate the differences in the AR(1)-RLS and the RRR methods' capabilities to detect outliers at the 2.5 σ level.

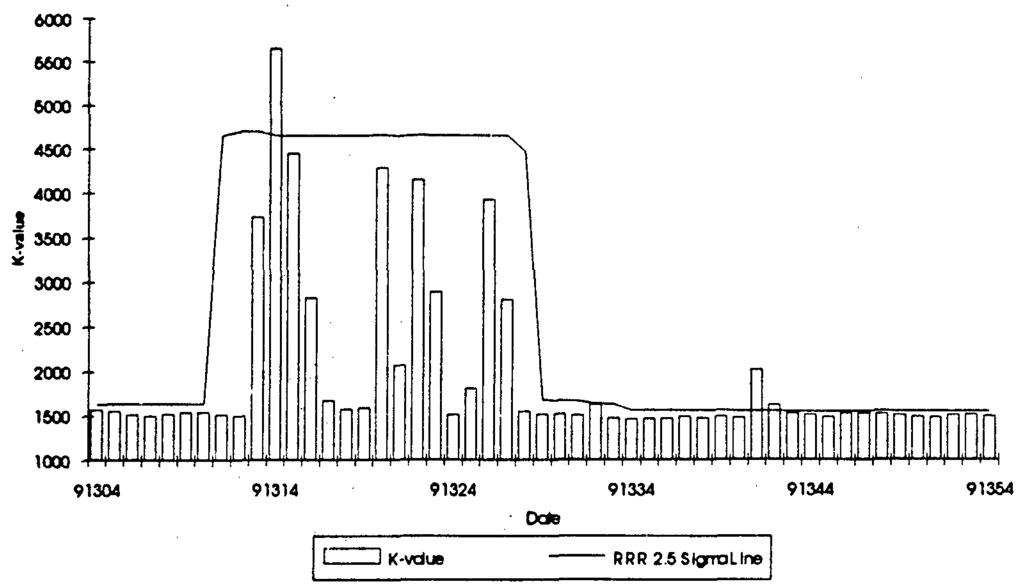

Figure 16. K-value with RRR 2.5 σ Line

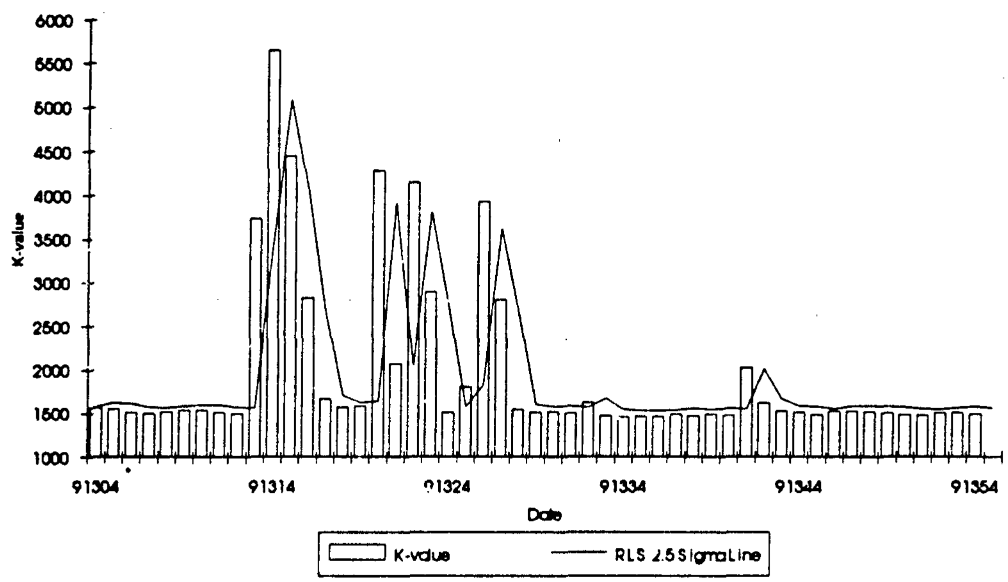

Figure 17. K-value with AR(1)-RLS 2.5 σ Line

The most striking feature of both of the RRR plots, Figures 16 and 17, is that the significance line tends to follow the data plot, anticipating when the k-values are going to rise. The problem here is that the RRR method uses the data after a point, as well as that before to predict the point. As in the case of the data presented here, the RRR method often overlooks obvious outliers in the data. The AR(1)-RLS model however, accurately predicts the major changes in the data. It illustrates the capability of the method to detect the significant outliers. While the 3.0 and 2.5 σ RRR methods only identified two and three outliers respectively, the AR(1)-RLS identified ten obvious outliers. The RRR methods only identified the most obvious and largest outlier.

Figure 18 is an enlargement of Figure 17. This figure more clearly illustrates how the AR(1)-RLS method fits the data. The leading edge is accurately identified as an outlier, but the trailing edge values are predicted by the AR(1) model as usual return to background level. The values for days 91320, 91322, and 91326 are high and identified as such. However, while days 92321, 91323, and 91327 are high, they represent the subsequent decay of the previous days large value and are accurately accounted for by the model. The observation that they are lower than the model predicts suggests that the days identified as outliers we very significant for this site. The simplicity is that AR(1)-RLS flags these values for further consideration by the AFTAC analyst while RRR misses them completely.

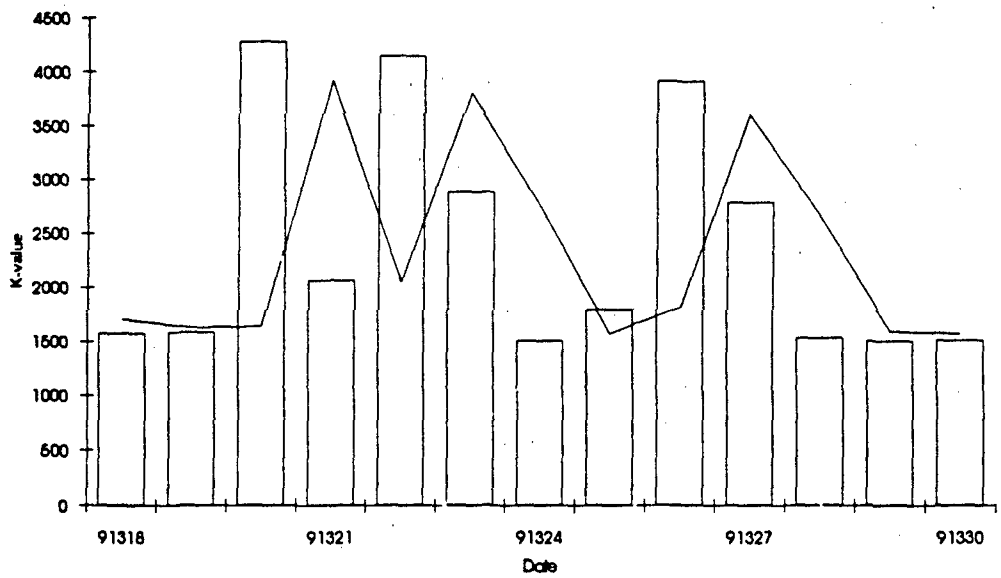

Figure 18. Enlargement of Figure 17

Subset Analysis

The final step in the method analysis process is to study the effect of different sizes of data sets on the detection of outliers. This study is necessary in order to determine the optimum sample size on which to perform the analysis.

After graphing the data, analysts are often interested in the statistics surrounding a particular point of interest. The question arises, is it an outlier or is it a good data point? This particular study was performed to look at the effect of population or window sizes on the outlier detection capability of the model.

Three sample sets with suspected outliers were selected. For each of these sample sets, subsets of 30, 50, 75, 100, and 300 days were used. Each data point that appeared to be an outlier, was identified as such in the analysis of every subset in which it occurred, regardless of the size of the subset. Where the results varied was on days that are very close to the threshold level of significance. That is to say, a data point that fell just below the 2.5-σ line of significance in one size subset might be above this cutoff in another. The determining factor appears to be the amount of noise in the data. In general, one would expect that the larger the sample size, the larger the number of outliers the model will detect. This was not necessarily the case here. Overall, the method was insensitive to the sample size. However, on the basis of work by Box and Jenkins, the minimum sample size should be 50 and preferably 100 should be used (Box and Jenkins, 1976: 18).

Summary

The autoregressive order one reweighted least squares (AR(1)-RLS) model produced the best results over the range of the data analyzed. The confidence tests, including the P-value for the individual coefficients, R^2, and the scale estimate each gave inconclusive results as to which coefficients should be kept in the model. In the final analysis, the AR(1)-RLS model was selected based upon the results of the stepwise regression. The stepwise regression, for each of the five data sets, indicated that lag one was the only significant predictor, and that lags two and three should not be used in the model.

The subset analysis performed, along with previous work by Box and Jenkins, suggests that a large data set size is desirable. The exact size of the data set to be used was not determined. There is no need to seek an exact best data set size, since the identification of significant outliers is very insensitive to the data set size. Data sets of 50 to 100 days (2 to 3 months) seem appropriate.

VI. Conclusions & Recommendations

Introduction

The objective of this research was to develop a methodology to improve on or supplant the existing procedure for identifying significant outliers in time series data. Comparisons were made between the results of the RRR method, the derivative method and the autoregressive order one reweighted least squares (AR(1)-RLS) method developed in this paper.

A summary, conclusions, and recommendations from this effort are presented based on the results of the techniques applied. On the basis of the work presented here, I concluded that the AR(1)-RLS method provided the best outlier detection model.

Observations

Based principally upon stepwise regression, the AR(1)-RLS can be expected to be an adequate, and probably optimal, model for fitting the full range of AFTAC data, regardless of the site. (The five sample data sets were selected by AFTAC to span this range and order one provided the best AR model for all sites)

For the five sample sets provided by AFTAC, AR(1)-RLS found every set of outliers that RRR found. Thus AR(1)-RLS can be expected to overlook few or none of the outliers that RRR can find.

AR(1)-RLS appears to be insensitive to the data set size used in performing the analysis. This is contrary to RRR which is extremely dependent on the data set size in determining whether a particular point is an outlier. Additionally, AR(1)-RLS does not require two weeks of data beyond the day of interest to perform its analysis, making it much more timely in detecting outliers.

Unlike the derivative method, AR(1)-RLS requires no special treatment of the data to handle missing data. Furthermore, no smoothing of the data is required to remove any non-stationarity.

Finally, AR(1)-RLS found four times as many outliers as RRR found in the data sets.

Conclusions

The AR(1)-RLS method is more effective than the RRR and should replace it. The inclusion of higher order lags is unjustified and the simplicity of AR(1) makes it an attractive method to use. The AR(1)-RLS method developed here is successful in locating all outliers identified by the RRR method as well as many others that the RRR method overlooks. Additionally, the AR(1)-RLS method will identify not-so-obvious outliers that bear further investigation by the analyst to determine their importance.

In addition to the obvious outliers that the AR(1)-RLS method detects, it does an excellent job at explaining why some data points, which appear to be outliers, are not outliers. By correctly fitting the data, a successive outlier is identified as a relic of the previous fluctuation and is not itself significant.

Recommendations

In order for this method to receive full acceptance by AFTAC, it will be necessary to modify Dr. Rousseeuw's PROGRESS code. The most important modification, and perhaps the only one absolutely necessary, is changing the σ cutoff value from a fixed 2.5 to an input variable. This will allow the PROGRESS code to operate at the same level of significance as AFTAC's recursive removal without regression method.

There are several other possibilities to improve upon the work presented in this thesis, the first of which is the addition of a spatial parameter to try to incorporate meteorological effects seen over geographically close sites. Consideration should be given to future improvements such as coupled space and time modeling for geographically related sampling sites.

Appendix A: Recursive Rejection without Regression BASIC Code

This appendix contains the BASIC code for AFTAC's recursive rejection without regression (RRR) method. This code was adapted from the PL/1 version of the code provided to me by AFTAC. The code was used to produce the RRR results discussed in Chapters IV and V.

```
'    RRR program
'
'   this program takes the value data and performs a Recursive Rejection
'   without Regression (RRR) on the data points. This is the method
'   currently in place at AFTAC/TNR. Code adapted from PL/1 code
provided
'   by AFTAC/TNR in November 1992.


ver$ = "RRR.bas, Version 1, written by Capt Keri L. Robinson, GNE93M,
8 Dec 92"
CLS
DEFINT I, L-N

TYPE file
        filename AS STRING * 40
END TYPE

DIM infile AS file
DIM outfile AS file
```

```
DIM tempfile AS file

INPUT "Enter the name of the data file including path:  "; infile.filename
OPEN infile.filename FOR INPUT AS #1

INPUT "Enter the name of the output file including path:  ";
outfile.filename

OPEN outfile.filename FOR OUTPUT AS #3

'    reading in data file for the number of records
n = 0
i = 0
DO
        n = n + 1
        INPUT #1, aa, bb
LOOP UNTIL (EOF(1))
CLOSE (1)

PRINT "This file contains "; n; " records."

REDIM a(1 TO 30)       '30 day array
REDIM value(1 TO n)     'daily value
REDIM jdate(1 TO n)    'Julian data of data
REDIM sd(1 TO n)        'daily standard deviation above background
```

```
REDIM sigout(1 TO n) AS STRING    'mark the status of the da   oint
REDIM smean(1 TO n)     'calculated background value
REDIM drop(1 TO n)      'number dropped from each window calcu. ation
REDIM dropped%(1 TO 30)

CONST False = 0
CONST True = NOT False

OPEN infile.filename FOR INPUT AS #1
FOR in = 1 TO n
        INPUT #1, jdate(in), value(in)
NEXT in
CLOSE (1)


FOR in = 15 TO n - 15
        npts = 0
        sum = 0
        sum2 = 0
        i = 0
        sigma = 3
        IF value(in) <> 0 THEN
                FOR ia = in - 14 TO in + 15
                        i = i + 1
                        a(i) = value(ia)
                NEXT ia
```

```
i = 0
DO
        i = i + 1
        a = a(i)
        dropped%(i) = False
        IF a <> 0 THEN
                sum = sum + a
                sum2 = sum2 + a * a
                npts = npts + 1
        END IF
LOOP UNTIL i = 30
amean = sum / npts
sdev = SQR((sum2 - (sum ^ 2 / npts)) / (npts - 1))
numptsdrp = 0
drp = 0
DO
        numptsdrp = drp
        drp = 0
        i = 0
        DO
                i = i + 1
                a = a(i)
                IF a <> 0 THEN
                        IF ABS((a - amean) / sdev) > sigma AND
dropped%(i) = False THEN
                                sum = sum - a
```

```
                    sum2 = sum2 - (a * a)
                    npts = npts - 1
                    drp = drp + 1
                    dropped%(i) = True
                ELSEIF dropped%(i) = True THEN
                    drp = drp + 1
                END IF
            ELSE
                drp = drp + 1
            END IF
        LOOP UNTIL i = 30
        sdev = 0
        sdev = SQR((sum2 - (sum ^ 2 / npts)) / (npts - 1))
        amean = sum / npts
    LOOP UNTIL numptsdrp = drp OR drp > 15 'OR npts < 15

    sd(in) = (value(in) - amean) / sdev
    smean(in) = amean
    drop(in) = drp
  ' PRINT jdate(in), value(in), smean(in), sd(in)
    IF sd(in) > sigma THEN
        sigout(in) = "+"     'Specifies the value as an outlier
'  ELSEIF sd(in) < sigma THEN
'     sigout(in) = "0"     'Specifies the values is a good data point
    END IF
ELSEIF value(in) = 0 THEN
```

```
            sigout(in) = "-"          'Specifies the value not used in
calculations
        END IF

NEXT in

FOR in = 1 TO n
        PRINT #3, jdate(in), value(in), USING "#####.#     "; smean(in);
        PRINT #3, USING "##.###     "; sd(in);
        PRINT #3, sigout(in), drop(in)
NEXT in


CLOSE (1)
CLOSE (3)
END
```

Appendix B: Derivative Method BASIC Code

This appendix contains the BASIC version of the derivative method written by Dr. Lloyd Currie. The code was adapted from the FORTRAN version provided by AFTAC. The results of this code are discussed in Chapters IV and V.

```
DECLARE SUB NormalStdDev (sampavg!(), j!, ndays%, sigma!(), stddev!())
DECLARE SUB interp (samp!(), n%)
DECLARE SUB average (samp!(), sampavg!(), n%)
'   program currie
'
'  this program takes the sample data and performs the currie
'  algorithm on the data. This is a modification of the FORTRAN version
'  of the currie code provided by AFTAC.


ver$ = "Currie.bas, Version 1, written by Capt Keri L. Robinson,
GNE93M, 1 Oct 92"
CLEAR

DEFINT I, L-N

TYPE file
      filename AS STRING * 40
END TYPE
```

```
DIM infile AS file
DIM outfile AS file
DIM tempfile AS file
.


INPUT "Enter the name of the data file including path:  "; infile.filename
OPEN infile.filename FOR INPUT AS #1

INPUT "Enter the name of the output file including path:  ";
outfile.filename

OPEN "d:\tmp\temp.out" FOR OUTPUT AS #3

'    reading in data file for the number of records
n = 0
i = 0
DO
        n = n + 1
        INPUT #1, a, b
LOOP UNTIL (EOF(1))
CLOSE (1)

PRINT "This file contains "; n, " records."

REDIM samp(1 TO n)      'Raw data
```

```
REDIM sampavg(1 TO n)  'Three day averaged data (smoothed)
REDIM dif3(1 TO n)      'First difference of smoothed data
REDIM dif33(1 TO n)     'Second difference of smoothed data
REDIM zfac1(1 TO n)     'Z-Factor of the first difference
REDIM zfac2(1 TO n)     'Z-Factor of the second difference
REDIM jdate(1 TO n)     'Julian data of data

OPEN infile.filename FOR INPUT AS #1

FOR in = 1 TO n
      INPUT #1, jdate(in), samp(in)
NEXT in
CLOSE (1)

'   data check and interpolating missing values up to 2 days
CALL interp(samp(), n)

'   calculating the 3 day moving average

CALL average(samp(), sampavg(), n)

'   calculating the first divided difference
'   This is an attempt at an unbiased first derivative by using points
'   in the numerical approximation to the derivative at a point which
```

```
' were not used in calculating sampavg. sampavg(i) is a function of (i-
1,i,i+1),
' so to do the first difference unbiased, you must go to sampavg(i-3) for
' backward difference in order not to use points used in sampavg(i)
' This all assumes that no data are missing

FOR i = 5 TO n - 1
    IF ((sampavg(i - 3) <> 0!) AND (sampavg(i) <> 0!)) THEN
        dif3(i) = sampavg(i) - sampavg(i - 3)
    END IF
NEXT

' calculating the second difference
' the method of using unbiased data applies here, but the method used
for
' the second difference is works out to be
' sampavg(i)"=sampavg(i+3)-2*sampavg(i)+sampavg(i-3)

FOR i = 5 TO n - 4
    IF (dif3(i + 3) <> 0! AND dif3(i) <> 0!) THEN
        dif33(i) = dif3(i + 3) - dif3(i)
    END IF
NEXT


' calculating sigma for dif3 and dif33
```

```
' here rbar is the range between the two values calculated above. This
' is used later to approximate sigma for the group of data of interest.
' This method is fully described in Thomas P. Ryans book "Statistical
' Methods for Quality Improvement" pp 82-86.

DO
    INPUT "How many days do you want in the window? (min 20)",
ndays
LOOP UNTIL ndays >= 20

ndays = INT(ndays / 2)
OPEN outfile.filename FOR OUTPUT AS #2


FOR j = (ndays + 2) TO (n - ndays)   'Needed to have sufficient days in the
i loop
    rbar1 = 0!
    rbar2 = 0!
    num1 = 0
    num2 = 0
    FOR i = j -- ndays TO j + ndays     'using a nday moving average
        IF ((dif3(i - 1) <> 0!) AND (dif3(i) <> 0!)) THEN
            rbar1 = ABS(dif3(i - 1) - dif3(i)) + rbar1
            num1 = num1 + 1
        END IF
        IF ((dif33(i - 1) <> 0!) AND (dif33(i) <> 0!)) THEN
```

```
            rbar2 = ABS(dif33(i - 1) - dif33(i)) + rbar2
            num2 = num2 + 1
        END IF
    NEXT i

    REDIM sigma(1 TO j)
    REDIM stddev(1 TO j)
    CALL NormalStdDev(sampavg(), j, ndays, sigma(), stddev())
'   PRINT #2, jdate(j), stddev(j)
'   Calculate sigma for the first and second difference.
'   The number 1.128 comes from a table constructed to allow the
average
'   of the ranges to be divided by this constant so that the resultant
'   number is an unbiased estimator of sigma.  This is from Table E, pg
434
'   of Ryans book.

    IF num1 = 0 THEN          'check for no data is calculation
        sigma1 = -1
    ELSE
        sigma1 = rbar1 / (1.128 * (num1))
    END IF
    IF num2 = 0 THEN          'check for no data is calculation
        sigma2 = -1
    ELSE
        sigma2 = rbar2 / (1.128 * (num2))
```

```
END IF
PRINT #3, rbar1, sigma1, rbar2, sigma2



' calculating zfactors for both dif3 and dif33
' The Z-factor or Z-score calculated below is a probability that a value
' is outside a range defined by a normal distribution.  Z-scores
represent
' the area under a normal curve from the mean to a pont on the curve.
' This assumes the value we want to compare to, mu, is zero.  The
' Z-score calculated here is (average-mu)/sigma.  The differences are
our
' average, and the estimated sigma is calculated above.

   IF num1 < 15 OR sigma1 = -1 THEN    'signifies not enough data
for good
        zfac1(j) = -1                  'statistics
   ELSE
        zfac1(j) = dif3(j) / sigma1
   END IF
   IF num2 < 15 OR sigma2 = -1 THEN
        zfac2(j) = -1
   ELSE
        zfac2(j) = dif33(j) / sigma2
```

```
    END IF

NEXT j

'   printing output


PRINT #2, "The following is based on an"; ndays * 2; "day moving
window"
PRINT #2,
PRINT #2, " Date      Sample     AVG3    DIF3   DIF33  ZFAC1
ZFAC2"
FOR i = 1 TO n
    PRINT #2, USING "#####    "; jdate(i);
    PRINT #2, USING "#####.##   "; samp(i); sampavg(i);
    PRINT #2, USING "####.###   "; dif3(i); dif33(i);
    PRINT #2, USING "###.####   "; zfac1(i); zfac2(i)
NEXT i

CLOSE (2)
CLOSE (3)

END

SUB average (samp(), sampavg(), n)
```

```
FOR i = 2 TO n - 1
        IF ((samp(i - 1) <> 0!) AND (samp(i + 1) <> 0!) AND (samp(i) <> 0!))
THEN
                sampavg(i) = (samp(i - 1) + samp(i) + samp(i + 1)) / 3!
        END IF
NEXT

END SUB

SUB interp (samp(), n)
'  data check and interpolating missing values up to 2 days
FOR i = 2 TO n - 2
        IF (samp(i) = 0! AND samp(i + 1) = 0! AND samp(i + 2) = 0!) THEN
                samp(i) = 0!
                samp(i + 1) = 0!
                samp(i + 2) = 0!
        ELSEIF (samp(i - 1) = 0! AND samp(i) = 0!) THEN
                samp(i) = 0!
        ELSEIF (samp(i) = 0! AND samp(i + 1) = 0!) THEN
                samp(i) = (samp(i - 1) * 2! + samp(i + 2)) / 3!
                samp(i + 1) = (samp(i - 1) + samp(i + 2) * 2!) / 3!
        ELSEIF (samp(i) = 0!) THEN
                samp(i) = (samp(i - 1) + samp(i + 1)) / 2!
        END IF
NEXT i
```

```
END SUB

SUB NormalStdDev (sampavg(), j, ndays, sigma(), stddev())

sampsqr = 0
samptot = 0
npts = 0

FOR ij = (j - ndays) TO (j + ndays)
        IF sampavg(ij) <> 0 THEN
                sampsqr = sampsqr + (sampavg(ij)) ^ 2
                samptot = samptot + sampavg(ij)
                npts = npts + 1
        END IF
NEXT ij
'calculate the sigma for the window
IF npts < 15 THEN                'min pts to be used in a sigma calculation
        sigma(j) = 0
ELSE
        sigma(j) = SQR((sampsqr - (samptot ^ 2 / npts)) / (npts - 1))
        bkg = samptot / npts
        stddev(j) = (sampavg(j) - bkg) / sigma(j)
END IF
'PRINT "The value is "; stddev(j); "outside the normal background."
END SUB
```

Appendix C: Data Listing for PROGRESS Run in Appendix D

This appendix contains a listing of the data used in the Test Methodology chapter of the thesis. This data was used as input for the PROGRESS code output in Appendix D.

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91201	1528.8	1543.5	1483.5	1420.8
91202	1473.8	1528.8	1543.5	1483.5
91203	1422.7	1473.8	1528.8	1543.5
91204	1389.9	1422.7	1473.8	1528.8
91205	1382	1389.9	1422.7	1473.8
91206	1406.2	1382	1389.9	1422.7
91207	1384.5	1406.2	1382	1389.9
91208	1445.8	1384.5	1406.2	1382
91209	1411.2	1445.8	1384.5	1406.2
91210	1383	1411.2	1445.8	1384.5
91211	1372.5	1383	1411.2	1445.8
91212	1369.7	1372.5	1383	1411.2
91213	1374.6	1369.7	1372.5	1383
91214	1436.5	1374.6	1369.7	1372.5
91215	1399.3	1436.5	1374.6	1369.7
91216	0	1399.3	1436.5	1374.6
91217	0	0	1399.3	1436.5

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91218	1421.7	0	0	1399.3
91219	1421.9	1421.7	0	0
91220	1422.7	1421.9	1421.7	0
91221	1427	1422.7	1421.9	1421.7
91222	1436	1427	1422.7	1421.9
91223	1440.2	1436	1427	1422.7
91224	1450.9	1440.2	1436	1427
91225	1463	1450.9	1440.2	1436
91226	1473.9	1463	1450.9	1440.2
91227	1506.9	1473.9	1463	1450.9
91228	1526	1506.9	1473.9	1463
91229	1517.3	1526	1506.9	1473.9
91230	1491.7	1517.3	1526	1506.9
91231	1429.5	1491.7	1517.3	1526
91232	1402.3	1429.5	1491.7	1517.3
91233	1383	1402.3	1429.5	1491.7
91234	1393.9	1383	1402.3	1429.5
91235	1398.1	1393.9	1383	1402.3
91236	1488	1398.1	1393.9	1383
91237	1555.6	1488	1398.1	1393.9
91238	1525.2	1555.6	1488	1398.1
91239	1614.3	1525.2	1555.6	1488
91240	1613.2	1614.3	1525.2	1555.6

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91241	1579	1613.2	1614.3	1525.2
91242	1358.9	1579	1613.2	1614.3
91243	1411.8	1358.9	1579	1613.2
91244	1479	1411.8	1358.9	1579
91245	1508.8	1479	1411.8	1358.9
91246	1461.7	1508.8	1479	1411.8
91247	1474.8	1461.7	1508.8	1479
91248	1460.1	1474.8	1461.7	1508.8
91249	1490.5	1460.1	1474.8	1461.7
91250	1448	1490.5	1460.1	1474.8
91251	1422.5	1448	1490.5	1460.1
91252	1412.7	1422.5	1448	1490.5
91253	1467.3	1412.7	1422.5	1448
91254	1497.2	1467.3	1412.7	1422.5
91255	1506.3	1497.2	1467.3	1412.7
91256	1547.3	1506.3	1497.2	1467.3
91257	1429.1	1547.3	1506.3	1497.2
91258	1512.3	1429.1	1547.3	1506.3
91259	1530.2	1512.3	1429.1	1547.3
91260	1529.8	1530.2	1512.3	1429.1
91261	1516.5	1529.8	1530.2	1512.3
91262	1422.4	1516.5	1529.8	1530.2
91263	1495.4	1422.4	1516.5	1529.8

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91264	1500.6	1495.4	1422.4	1516.5
91265	1488.6	1500.6	1495.4	1422.4
91266	1516.3	1488.6	1500.6	1495.4
91267	1512.5	1516.3	1488.6	1500.6
91268	1505	1512.5	1516.3	1488.6
91269	1497.7	1505	1512.5	1516.3
91270	1436.1	1497.7	1505	1512.5
91271	1457.2	1436.1	1497.7	1505
91272	1556.6	1457.2	1436.1	1497.7
91273	1546.6	1556.6	1457.2	1436.1
91274	1508.9	1546.6	1556.6	1457.2
91275	1478.5	1508.9	1546.6	1556.6
91276	1481	1478.5	1508.9	1546.6
91277	1462.8	1481	1478.5	1508.9
91278	1520.5	1462.8	1481	1478.5
91279	1532.2	1520.5	1462.8	1481
91280	1500.1	1532.2	1520.5	1462.8
91281	1493.6	1500.1	1532.2	1520.5
91282	1481.5	1493.6	1500.1	1532.2
91283	1501.5	1481.5	1493.6	1500.1
91284	1521.4	1501.5	1481.5	1493.6
91285	1467	1521.4	1501.5	1481.5
91286	1466.1	1467	1521.4	1501.5

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91287	1480.6	1466.1	1467	1521.4
91288	1544.1	1480.6	1466.1	1467
91289	1543.5	1544.1	1480.6	1466.1
91290	1539.9	1543.5	1544.1	1480.6
91291	1490.5	1539.9	1543.5	1544.1
91292	1495.5	1490.5	1539.9	1543.5
91293	1560.1	1495.5	1490.5	1539.9
91294	1559.4	1560.1	1495.5	1490.5
91295	1548.5	1559.4	1560.1	1495.5
91296	1584.7	1548.5	1559.4	1560.1
91297	1608.6	1584.7	1548.5	1559.4
91298	1624.5	1608.6	1584.7	1548.5
91299	0	1624.5	1608.6	1584.7
91300	1524.6	0	1624.5	1608.6
91301	1535	1524.6	0	1624.5
91302	1521.4	1535	1524.6	0
91303	1513.2	1521.4	1535	1524.6
91304	1569.7	1513.2	1521.4	1535
91305	1555.6	1569.7	1513.2	1521.4
91306	1512.3	1555.6	1569.7	1513.2
91307	1499.6	1512.3	1555.6	1569.7
91308	1521.4	1499.6	1512.3	1555.6
91309	1543.2	1521.4	1499.6	1512.3

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91310	1538.4	1543.2	1521.4	1499.6
91311	1511.5	1538.4	1543.2	1521.4
91312	1503	1511.5	1538.4	1543.2
91313	3735.5	1503	1511.5	1538.4
91314	5655.4	3735.5	1503	1511.5
91315	4450.3	5655.4	3735.5	1503
91316	2827	4450.3	5655.4	3735.5
91317	1667.7	2827	4450.3	5655.4
91318	1576.1	1667.7	2827	4450.3
91319	1587	1576.1	1667.7	2827
91320	4287.1	1587	1576.1	1667.7
91321	2072	4287.1	1587	1576.1
91322	4156.3	2072	4287.1	1587
91323	2898.6	4156.3	2072	4287.1
91324	1513.8	2898.6	4156.3	2072
91325	1805.6	1513.8	2898.6	4156.3
91326	3933.3	1805.6	1513.8	2898.6
91327	2806.1	3933.3	1805.6	1513.8
91328	1545.6	2806.1	3933.3	1805.6
91329	1516.7	1545.6	2806.1	3933.3
91330	1521.5	1516.7	1545.6	2806.1
91331	1504.5	1521.5	1516.7	1545.6
91332	1627.5	1504.5	1521.5	1516.7

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91333	1475	1627.5	1504.5	1521.5
91334	1460	1475	1627.5	1504.5
91335	1463.4	1460	1475	1627.5
91336	1465.3	1463.4	1460	1475
91337	1493.2	1465.3	1463.4	1460
91338	1472.7	1493.2	1465.3	1463.4
91339	1493.7	1472.7	1493.2	1465.3
91340	1482.7	1493.7	1472.7	1493.2
91341	2023.8	1482.7	1493.7	1472.7
91342	1623.4	2023.8	1482.7	1493.7
91343	1527.9	1623.4	2023.8	1482.7
91344	1515.4	1527.9	1623.4	2023.8
91345	1485.1	1515.4	1527.9	1623.4
91346	1527.8	1485.1	1515.4	1527.9
91347	1527	1527.8	1485.1	1515.4
91348	1520.2	1527	1527.8	1485.1
91349	1510	1520.2	1527	1527.8
91350	1484.7	1510	1520.2	1527
91351	1478.8	1484.7	1510	1520.2
91352	1503.8	1478.8	1484.7	1510
91353	1516	1503.8	1478.8	1484.7
91354	1495.5	1516	1503.8	1478.8
91355	1489.6	1495.5	1516	1503.8

Julian Date	K-Value	Lag 1	Lag 2	Lag 3
91356	1488.2	1489.6	1495.5	1516
91357	1480.6	1488.2	1489.6	1495.5
91358	1463.1	1480.6	1488.2	1489.6
91359	0	1463.1	1480.6	1488.2
91360	1507.1	0	1463.1	1480.6
91361	1506.5	1507.1	0	1463.1
91362	1522.6	1506.5	1507.1	0
91363	1534.8	1522.6	1506.5	1507.1
91364	1488	1534.8	1522.6	1506.5
91365	1483	1488	1534.8	1522.6

Appendix D: Sample Output from PROGRESS Code

This appendix contains the output from the PROGRESS code provided by Dr. Rousseeuw. The input data are listed in Appendix C. This output is discussed in the Test Methodology Chapter.

```
                         *******************
                         * P R O G R E S S *
                         *******************


 This program performs a robust regression analysis based on
 the least median of squares (LMS) method as described in
    P. Rousseeuw (1984), Least Median of Squares Regression,
    Journal of the American Statistical Association, 79, 871-880.
 A user manual to this program is the book:
    P. Rousseeuw and A. Leroy (1987), Robust Regression
    and Outlier Detection, Wiley, New York.

 DATA SET = DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

 REGRESSION WITH A CONSTANT TERM.

 NUMBER OF CASES         =    165
 NUMBER OF COEFFICIENTS (INCLUDING CONSTANT TERM) =      2

 THE EXTENSIVE SEARCH VERSION WILL BE USED.

 TREATMENT OF MISSING VALUES IN OPTION 1: THIS MEANS THAT A CASE WITH A
 MISSING VALUE FOR AT LEAST ONE VARIABLE WILL BE DELETED.

  LARGE OUTPUT IS WANTED.

 YOUR DATA RESIDE IN FILE        : 201_365.DAT

 VARIABLE LAG1 VALUE HAS A MISSING VALUE FOR    4 CASES.
 VARIABLE      KVALUE HAS A MISSING VALUE FOR    4 CASES.

 CASE HAS A MISSING VALUE FOR VARIABLES (VARIABLE NUMBER    3 IS THE
RESPONSE)
 ----                          ---------
   16                           3
   17                           1    3
   18                           1
   99                           3
  100                           1
```

159	3
160	1

THERE ARE 158 CASES STAYING IN THE ANALYSIS.

THE OBSERVATIONS, AFTER TREATMENT OF MISSING VALUES ARE:

	LAG1 VALUE	KVALUE
1	1543.5000	1528.8000
2	1528.8000	1473.8000
3	1473.8000	1422.7000
4	1422.7000	1389.9000
5	1389.9000	1382.0000
6	1382.0000	1406.2000
7	1406.2000	1384.5000
8	1384.5000	1445.8000
9	1445.8000	1411.2000
10	1411.2000	1383.0000
11	1383.0000	1372.5000
12	1372.5000	1369.7000
13	1369.7000	1374.6000
14	1374.6000	1436.5000
15	1436.5000	1399.3000
19	1421.7000	1421.9000
20	1421.9000	1422.7000
21	1422.7000	1427.0000
22	1427.0000	1436.0000
23	1436.0000	1440.2000
24	1440.2000	1450.9000
25	1450.9000	1463.0000
26	1463.0000	1473.9000
27	1473.9000	1506.9000
28	1506.9000	1526.0000
29	1526.0000	1517.3000
30	1517.3000	1491.7000
31	1491.7000	1429.5000
32	1429.5000	1402.3000
33	1402.3000	1383.0000
34	1383.0000	1393.9000
35	1393.9000	1398.1000
36	1398.1000	1488.0000
37	1488.0000	1555.6000
38	1555.6000	1525.2000
39	1525.2000	1614.3000
40	1614.3000	1613.2000
41	1613.2000	1579.0000
42	1579.0000	1358.9000
43	1358.9000	1411.8000
44	1411.8000	1479.0000
45	1479.0000	1508.8000
46	1508.8000	1461.7000
47	1461.7000	1474.8000

48	1474.8000	1460.1000
49	1460.1000	1490.5000
50	1490.5000	1448.0000
51	1448.0000	1422.5000
52	1422.5000	1412.7000
53	1412.7000	1467.3000
54	1467.3000	1497.2000
55	1497.2000	1506.3000
56	1506.3000	1547.3000
57	1547.3000	1429.1000
58	1429.1000	1512.3000
59	1512.3000	1530.2000
60	1530.2000	1529.8000
61	1529.8000	1516.5000
62	1516.5000	1422.4000
63	1422.4000	1495.4000
64	1495.4000	1500.6000
65	1500.6000	1488.6000
66	1488.6000	1516.3000
67	1516.3000	1512.5000
68	1512.5000	1505.0000
69	1505.0000	1497.7000
70	1497.7000	1436.1000
71	1436.1000	1457.2000
72	1457.2000	1556.6000
73	1556.6000	1546.6000
74	1546.6000	1508.9000
75	1508.9000	1478.5000
76	1478.5000	1481.0000
77	1481.0000	1462.8000
78	1462.8000	1520.5000
79	1520.5000	1532.2000
80	1532.2000	1500.1000
81	1500.1000	1493.6000
82	1493.6000	1481.5000
83	1481.5000	1501.5000
84	1501.5000	1521.4000
85	1521.4000	1467.0000
86	1467.0000	1466.1000
87	1466.1000	1480.6000
88	1480.6000	1544.1000
89	1544.1000	1543.5000
90	1543.5000	1539.9000
91	1539.9000	1490.5000
92	1490.5000	1495.5000
93	1495.5000	1560.1000
94	1560.1000	1559.4000
95	1559.4000	1548.5000
96	1548.5000	1584.7000
97	1584.7000	1608.6000
98	1608.6000	1624.5000
101	1524.6000	1535.0000

102	1535.0000	1521.4000
103	1521.4000	1513.2000
104	1513.2000	1569.7000
105	1569.7000	1555.6000
106	1555.6000	1512.3000
107	1512.3000	1499.6000
108	1499.6000	1521.4000
109	1521.4000	1543.2000
110	1543.2000	1538.4000
111	1538.4000	1511.5000
112	1511.5000	1503.0000
113	1503.0000	3735.5000
114	3735.5000	5655.4000
115	5655.4000	4450.3000
116	4450.3000	2827.0000
117	2827.0000	1667.7000
118	1667.7000	1576.1000
119	1576.1000	1587.0000
120	1587.0000	4287.1000
121	4287.1000	2072.0000
122	2072.0000	4156.3000
123	4156.3000	2898.6000
124	2898.6000	1513.8000
125	1513.8000	1805.6000
126	1805.6000	3933.3000
127	3933.3000	2806.1000
128	2806.1000	1545.6000
129	1545.6000	1516.7000
130	1516.7000	1521.5000
131	1521.5000	1504.5000
132	1504.5000	1627.5000
133	1627.5000	1475.0000
134	1475.0000	1460.0000
135	1460.0000	1463.4000
136	1463.4000	1465.3000
137	1465.3000	1493.2000
138	1493.2000	1472.7000
139	1472.7000	1493.7000
140	1493.7000	1482.7000
141	1482.7000	2023.8000
142	2023.8000	1623.4000
143	1623.4000	1527.9000
144	1527.9000	1515.4000
145	1515.4000	1485.1000
146	1485.1000	1527.8000
147	1527.8000	1527.0000
148	1527.0000	1520.2000
149	1520.2000	1510.0000
150	1510.0000	1484.7000
151	1484.7000	1478.8000
152	1478.8000	1503.8000
153	1503.8000	1516.0000

154	1516.0000	1495.5000
155	1495.5000	1489.6000
156	1489.6000	1488.2000
157	1488.2000	1480.6000
158	1480.6000	1463.1000
161	1507.1000	1506.5000
162	1506.5000	1522.6000
163	1522.6000	1534.8000
164	1534.8000	1488.0000
165	1488.0000	1483.0000

DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

```
OBSERVED
   KVALUE I-+----+----+----+----+----+----+----+----+----+----+-I
          I                                                     I
.5655E+04 +                            1                        +
          I                                                     I
          I                                                     I
          I                                                     I
          I                                                     I
          +                                                     +
          I                                                     I
          I                                                     I
          I                                                     I
          I                                                     I
          +                                                  1  +
          I   1                                                 I
          I         1                                           I
          I                                                     I
          I                                                     I
          +      1                                              +
          I  1                                                  I
          I                                                     I
          I                                                     I
          I                                                     I
          +                                                     +
          I                                                     I
          I                                                     I
          I                                                     I
          I                                 1                   I
          +                              1     1                +
          I                                                     I
          I                                                     I
          I                                                     I
          I                                                     I
          +                                                     +
          I                                                     I
          I  1                                1                 I
          I                                                     I
          I  1                                                  I
          +                                                     +
          I  261   1         1                                  I
          I 5**2            11                                  I
          I **1                                                 I
          I                                                     I
.1359E+04 +   1                                                 +
          I                                                     I
          I-+----+----+----+----+----+----+----+----+----+----+-I
    .1359E+04                                                .5655E+04

                                              OBSERVED LAG1 VALUE
```

MEDIANS =

LAG1 VALUE	KVALUE
1503.4000	1501.0500

DISPERSIONS =

LAG1 VALUE	KVALUE
53.7442	52.4100

THE STANDARDIZED OBSERVATIONS ARE:

	LAG1 VALUE	KVALUE
1	.7461	.5295
2	.4726	-.5199
3	-.5508	-1.4949
4	-1.5016	-2.1208
5	-2.1119	-2.2715
6	-2.2588	-1.8098
7	-1.8086	-2.2238
8	-2.2123	-1.0542
9	-1.0717	-1.7144
10	-1.7155	-2.2524
11	-2.2402	-2.4528
12	-2.4356	-2.5062
13	-2.4877	-2.4127
14	-2.3965	-1.2316
15	-1.2448	-1.9414
19	-1.5202	-1.5102
20	-1.5164	-1.4949
21	-1.5016	-1.4129
22	-1.4215	-1.2412
23	-1.2541	-1.1610
24	-1.1759	-.9569
25	-.9768	-.7260
26	-.7517	-.5180
27	-.5489	.1116
28	.0651	.4761
29	.4205	.3101
30	.2586	-.1784
31	-.2177	-1.3652
32	-1.3750	-1.8842
33	-1.8811	-2.2524
34	-2.2402	-2.0445
35	-2.0374	-1.9643
36	-1.9593	-.2490
37	-.2865	1.0408
38	.9713	.4608
39	.4056	2.1608
40	2.0635	2.1399

41	2.0430	1.4873
42	1.4067	-2.7123
43	-2.6887	-1.7025
44	-1.7044	-.4207
45	-.4540	.1479
46	.1005	-.7508
47	-.7759	-.5009
48	-.5321	-.7813
49	-.8057	-.2013
50	-.2400	-1.0122
51	-1.0308	-1.4988
52	-1.5053	-1.6858
53	-1.6876	-.6440
54	-.6717	-.0735
55	-.1154	.1002
56	.0540	.8825
57	.8168	-1.3728
58	-1.3825	.2147
59	.1656	.5562
60	.4987	.5486
61	.4912	.2948
62	.2437	-1.5007
63	-1.5071	-.1078
64	-.1489	-.0036
65	-.0521	-.2376
66	-.2754	.2910
67	.2400	.2185
68	.1693	.0754
69	.0298	-.0639
70	-.1061	-1.2393
71	-1.2522	-.8367
72	-.8596	1.0599
73	.9899	.8691
74	.8038	.1498
75	.1023	-.4303
76	-.4633	-.3826
77	-.4168	-.7298
78	-.7554	.3711
79	.3182	.[illegible]44
80	.5359	-.0181
81	-.0614	-.1421
82	-.1823	-.3730
83	-.4075	.0086
84	-.0354	.3883
85	.3349	-.6497
86	-.6773	-.6669
87	-.6940	-.3902
88	-.4242	.8214
89	.7573	.8100
90	.7461	.7413
91	.6791	-.2013
92	-.2400	-.1059

93	-.1470	1.1267
94	1.0550	1.1133
95	1.0420	.9054
96	.8392	1.5961
97	1.5127	2.0521
98	1.9574	2.3555
101	.3945	.6478
102	.5880	.3883
103	.3349	.2318
104	.1823	1.3099
105	1.2336	1.0408
106	.9713	.2147
107	.1656	-.0277
108	-.0707	.3883
109	.3349	.8042
110	.7405	.7127
111	.6512	.1994
112	.1507	.0372
113	-.0074	42.6341
114	41.5319	79.2664
115	77.2548	56.2727
116	54.8319	25.2996
117	24.6278	3.1797
118	3.0571	1.4320
119	1.3527	1.6400
120	1.5555	53.1588
121	51.7953	10.8939
122	10.5797	50.6631
123	49.3616	26.6657
124	25.9600	.2433
125	.1935	5.8109
126	5.6229	46.4082
127	45.2123	24.9008
128	24.2389	.8500
129	.7852	.2986
130	.2475	.3902
131	.3368	.0658
132	.0205	2.4127
133	2.3091	-.4970
134	-.5284	-.7832
135	-.8075	-.7184
136	-.7443	-.6821
137	-.7089	-.1498
138	-.1898	-.5409
139	-.5712	-.1402
140	-.1805	-.3501
141	-.3852	9.9742
142	9.6829	2.3345
143	2.2328	.5123
144	.4559	.2738
145	.2233	-.3043
146	-.3405	.5104

147	.4540	.4951
148	.4391	.3654
149	.3126	.1708
150	.1228	-.3120
151	-.3479	-.4245
152	-.4577	.0525
153	.0074	.2853
154	.2344	-.1059
155	-.1470	-.2185
156	-.2568	-.2452
157	-.2828	-.3902
158	-.4242	-.7241
161	.0688	.1040
162	.0577	.4112
163	.3572	.6440
164	.5842	-.2490
165	-.2865	-.3444

PEARSON CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
(KVALUE IS THE RESPONSE VARIABLE)

LAG1 VALUE	1.00	
KVALUE	.62	1.00

SPEARMAN RANK CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
(KVALUE IS THE RESPONSE VARIABLE)

LAG1 VALUE	1.00	
KVALUE	.73	1.00

****

LEAST SQUARES REGRESSION

****

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE	P - VALUE
LAG1 VALUE	.61790	.06297	9.81196	.00000
CONSTANT	625.14990	109.59590	5.70413	.00000

SUM OF SQUARES = 34251960.00000

DEGREES OF FREEDOM = 156

SCALE ESTIMATE = 468.57640

VARIANCE - COVARIANCE MATRIX =

```
 .3966D-02
-.6490D+01      .1201D+05
```

COEFFICIENT OF DETERMINATION (R SQUARED) = .38163

THE F-VALUE = 96.275 (WITH 1 AND 156 DF) P - VALUE = .00000

OBSERVED KVALUE	ESTIMATED KVALUE	RESIDUAL	NO	RES/SC
1528.80000	1578.88500	-50.08472	1	-.11
1473.80000	1569.80200	-96.00171	2	-.20
1422.70000	1535.81700	-113.11690	3	-.24
1389.90000	1504.24200	-114.34190	4	-.24
1382.00000	1483.97500	-101.97490	5	-.22
1406.20000	1479.09300	-72.89331	6	-.16
1384.50000	1494.04700	-109.54660	7	-.23
1445.80000	1480.63800	-34.83801	8	-.07
1411.20000	1518.51600	-107.31570	9	-.23
1383.00000	1497.13600	-114.13610	10	-.24
1372.50000	1479.71100	-107.21120	11	-.23
1369.70000	1473.22300	-103.52320	12	-.22
1374.60000	1471.49300	-96.89307	13	-.21
1436.50000	1474.52100	-38.02075	14	-.08
1399.30000	1512.76900	-113.46900	15	-.24
1421.90000	1503.62400	-81.72400	19	-.17
1422.70000	1503.74800	-81.04773	20	-.17
1427.00000	1504.24200	-77.24194	21	-.16
1436.00000	1506.89900	-70.89893	22	-.15
1440.20000	1512.46000	-72.26025	23	-.15
1450.90000	1515.05500	-64.15527	24	-.14
1463.00000	1521.66700	-58.66699	25	-.13
1473.90000	1529.14400	-55.24353	26	-.12
1506.90000	1535.87900	-28.97864	27	-.06
1526.00000	1556.27000	-30.26953	28	-.06
1517.30000	1568.07200	-50.77148	29	-.11
1491.70000	1562.69600	-70.99585	30	-.15
1429.50000	1546.87700	-117.37740	31	-.25
1402.30000	1508.44400	-106.14380	32	-.23
1383.00000	1491.63700	-108.63670	33	-.23
1393.90000	1479.71100	-85.81116	34	-.18
1398.10000	1486.44600	-88.34644	35	-.19
1488.00000	1489.04200	-1.04150	36	.00
1555.60000	1544.59100	11.00891	37	.02
1525.20000	1586.36100	-61.16150	38	-.13
1614.30000	1567.57700	46.72290	39	.10
1613.20000	1622.63200	-9.43250	40	-.02
1579.00000	1621.95300	-42.95264	41	-.09
1358.90000	1600.82000	-241.92040	42	-.52

1411.80000	1464.82000	-53.01978	43	-.11
1479.00000	1497.50700	-18.50684	44	-.04
1508.80000	1539.03000	-30.22998	45	-.06
1461.70000	1557.44400	-95.74365	46	-.20
1474.80000	1528.34000	-53.54028	47	-.11
1460.10000	1536.43500	-76.33484	48	-.16
1490.50000	1527.35200	-36.85156	49	-.08
1448.00000	1546.13600	-98.13599	50	-.21
1422.50000	1519.87500	-97.37500	51	-.21
1412.70000	1504.11800	-91.41846	52	-.20
1467.30000	1498.06300	-30.76294	53	-.07
1497.20000	1531.80100	-34.60059	54	-.07
1506.30000	1550.27600	-43.97583	55	-.09
1547.30000	1555.89900	-8.59875	56	-.02
1429.10000	1581.23300	-152.13290	57	-.32
1512.30000	1508.19700	4.10352	58	.01
1530.20000	1559.60600	-29.40625	59	-.06
1529.80000	1570.66700	-40.86670	60	-.09
1516.50000	1570.42000	-53.91956	61	-.12
1422.40000	1562.20100	-139.80140	62	-.30
1495.40000	1504.05700	-8.65662	63	-.02
1500.60000	1549.16400	-48.56360	64	-.10
1488.60000	1552.37700	-63.77673	65	-.14
1516.30000	1544.96200	-28.66187	66	-.06
1512.50000	1562.07800	-49.57788	67	-.11
1505.00000	1559.73000	-54.72974	68	-.12
1497.70000	1555.09500	-57.39551	69	-.12
1436.10000	1550.58500	-114.48470	70	-.24
1457.20000	1512.52200	-55.32202	71	-.12
1556.60000	1525.56000	31.04028	72	.07
1546.60000	1586.97900	-40.37939	73	-.09
1508.90000	1580.80000	-71.90027	74	-.15
1478.50000	1557.50500	-79.00537	75	-.17
1481.00000	1538.72100	-57.72107	76	-.12
1462.80000	1540.26600	-77.46582	77	-.17
1520.50000	1529.02000	-8.52002	78	-.02
1532.20000	1564.67300	-32.47314	79	-.07
1500.10000	1571.90200	-71.80249	80	-.15
1493.60000	1552.06800	-58.46777	81	-.12
1481.50000	1548.05100	-66.55139	82	-.14
1501.50000	1540.57500	-39.07471	83	-.08
1521.40000	1552.93300	-31.53284	84	-.07
1467.00000	1565.22900	-98.22925	85	-.21
1466.10000	1531.61500	-65.51526	86	-.14
1480.60000	1531.05900	-50.45911	87	-.11
1544.10000	1540.01900	4.08130	88	.01
1543.50000	1579.25600	-35.75562	89	-.08
1539.90000	1578.88500	-38.98474	90	-.08
1490.50000	1576.66000	-86.16040	91	-.18
1495.50000	1546.13600	-50.63599	92	-.11
1560.10000	1549.22500	10.87451	93	.02
1559.40000	1589.14200	-29.74207	94	-.06

1548.50000	1588.70900	-40.20947	95	-.09
1584.70000	1581.97400	2.72559	96	.01
1608.60000	1604.34300	4.25745	97	.01
1624.50000	1619.11000	5.38965	98	.01
1535.00000	1567.20600	-32.20642	101	-.07
1521.40000	1573.63300	-52.23254	102	-.11
1513.20000	1565.22900	-52.02930	103	-.11
1569.70000	1560.16200	9.53760	104	.02
1555.60000	1595.07400	-39.47400	105	-.08
1512.30000	1586.36100	-74.06140	106	-.16
1499.60000	1559.60600	-60.00623	107	-.13
1521.40000	1551.75900	-30.35876	108	-.06
1543.20000	1565.22900	-22.02930	109	-.05
1538.40000	1578.69900	-40.29944	110	-.09
1511.50000	1575.73400	-64.23352	111	-.14
1503.00000	1559.11200	-56.11182	112	-.12
3735.50000	1553.86000	2181.64000	113	4.66
5655.40000	2933.33100	2722.06900	114	5.81
4450.30000	4119.64500	330.65530	115	.71
2827.00000	3375.00800	-548.00830	116	-1.17
1667.70000	2371.96500	-704.26490	117	-1.50
1576.10000	1655.62900	-79.52856	118	-.17
1587.00000	1599.02900	-12.02856	119	-.03
4287.10000	1605.76400	2681.33600	120	5.72
2072.00000	3274.16700	-1202.16700	121	-2.57
4156.30000	1905.44700	2250.85300	122	4.80
2898.60000	3193.34400	-294.74440	123	-.63
1513.80000	2416.20700	-902.40650	124	-1.93
1805.60000	1560.53300	245.06690	125	.52
3933.30000	1740.83700	2192.46300	126	4.68
2806.10000	3055.55200	-249.45190	127	-.53
1545.60000	2359.05100	-813.45060	128	-1.74
1516.70000	1580.18200	-63.48242	129	-.14
1521.50000	1562.32500	-40.82495	130	-.09
1504.50000	1565.29100	-60.79102	131	-.13
1627.50000	1554.78700	72.71338	132	.16
1475.00000	1630.78900	-155.78880	133	-.33
1460.00000	1536.55800	-76.55835	134	-.16
1463.40000	1527.29000	-63.88977	135	-.14
1465.30000	1529.39100	-64.09070	136	-.14
1493.20000	1530.56500	-37.36475	137	-.08
1472.70000	1547.80400	-75.10425	138	-.16
1493.70000	1535.13700	-41.43726	139	-.09
1482.70000	1548.11300	-65.41321	140	-.14
2023.80000	1541.31600	482.48390	141	1.03
1623.40000	1875.66400	-252.26420	142	-.54
1527.90000	1628.25500	-100.35530	143	-.21
1515.40000	1569.24600	-53.84558	144	-.11
1485.10000	1561.52200	-76.42175	145	-.16
1527.80000	1542.79900	-14.99915	146	-.03
1527.00000	1569.18400	-42.18384	147	-.09
1520.20000	1568.68900	-48.48950	148	-.10

1510.00000	1564.48800	-54.48755	149	-.12
1484.70000	1558.18500	-73.48511	150	-.16
1478.80000	1542.55200	-63.75195	151	-.14
1503.80000	1538.90600	-35.10645	152	-.07
1516.00000	1554.35400	-38.35400	153	-.08
1495.50000	1561.89200	-66.39246	154	-.14
1489.60000	1549.22500	-59.62549	155	-.13
1488.20000	1545.58000	-57.37988	156	-.12
1480.60000	1544.71500	-64.11475	157	-.14
1463.10000	1540.01900	-76.91870	158	-.16
1506.50000	1556.39300	-49.89307	161	-.11
1522.60000	1556.02200	-33.42249	162	-.07
1534.80000	1565.97100	-31.17065	163	-.07
1488.00000	1573.50900	-85.50903	164	-.18
1483.00000	1544.59100	-61.59106	165	-.13

DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

--- L E A S T   S Q U A R E S ---

```
STAND. RESIDUAL I-+----+----+----+----+----+----+----+----+----+----+-I
                I                                                     I
      .5809E+01 +   1                        1                        +
                I                                                     I
                I                                                     I
                I         1                                           I
                I  1   1                                              I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
            2.5 I+++++++++++++++++++++++++++++++++++++++++++++++++++++I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                                                     +
                I  1                                                  I
                I                                                     I
                I                                                  1  I
                I  1                                                  I
                +                                                     +
                I  1                                                  I
            0.0 I-***2------------------------------------------------I
                I  111                                                I
                I   1    1                     1  1                   I
                +                                                     +
                I                                                     I
                I                                    1                I
                I                  1                                  I
                I                                                     I
                +                 11                                  +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
           -2.5 ++++++++++++++++++++++++++++++++++++1++++++++++++++++++
                I                                                     I
                I-+----+----+----+----+----+----+----+----+----+----+-I
         .1465E+04                                                .4120E+04

                                               ESTIMATED     KVALUE
```

DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

--- L E A S T   S Q U A R E S ---

```
STAND. RESIDUAL
           I-+----+----+----+----+----+----+----+----+----+----+-I
           I                                                     I
 .5809E+01 +                                   1 1               +
           I                                                     I
           I                                                     I
           I                                     1               I
           I                                   1   1             I
           +                                                     +
           I                                                     I
           I                                                     I
           I                                                     I
           I                                                     I
           +                                                     +
           I                                                     I
           I                                                     I
           I                                                     I
       2.5 I+++++++++++++++++++++++++++++++++++++++++++++++++++++I
           +                                                     +
           I                                                     I
           I                                                     I
           I                                                     I
           I                                                     I
           +                                                     +
           I                                           1         I
           I                                                     I
           I                                   1                 I
           I                                      1              I
           +                                                     +
           I                                        1            I
       0.0 I-4334123433432343332334333433422343-2---323233343131-I
           I                  11                     1           I
           I             1                        11   1         I
           +                                                     +
           I                                                     I
           I                                    1                I
           I                                    1                I
           I                                                     I
           +                                      11             +
           I                                                     I
           I                                                     I
           I                                                     I
           I                                                     I
      -2.5 ++++++++++++++++++++++++++++++++++++1++++++++++++++++++
           I                                                     I
           I-+----+----+----+----+----+----+----+----+----+----+-I
             1                                               165
                                 INDEX OF THE OBSERVATION
```

****
LEAST MEDIAN OF SQUARES REGRESSION
****

THE MINIMIZATION OF THE 80TH ORDERED SQUARED RESIDUAL IS PERFORMED.

ON A TOTAL OF 1001 SUBSAMPLES (OF 2 POINTS OUT OF 158)
1 SUBSAMPLES LED TO A SINGULAR SYSTEM OF EQUATIONS.
THE SOLUTION IS ONLY BASED ON THE GOOD SUBSAMPLES.

VARIABLE	COEFFICIENT
LAG1 VALUE	.97052
CONSTANT	45.42432

FINAL SCALE ESTIMATE = 29.29300

COEFFICIENT OF DETERMINATION = .69456

OBSERVED KVALUE	ESTIMATED KVALUE	RESIDUAL	NO	RES/SC
1528.80000	1543.41900	-14.61877	1	-.50
1473.80000	1529.15200	-55.35217	2	-1.89
1422.70000	1475.77400	-53.07385	3	-1.81
1389.90000	1426.18000	-36.28015	4	-1.24
1382.00000	1394.34700	-12.34729	5	-.42
1406.20000	1386.68000	19.51978	6	.67
1384.50000	1410.16700	-25.66663	7	-.88
1445.80000	1389.10600	56.69360	8	1.94
1411.20000	1448.59900	-37.39929	9	-1.28
1383.00000	1415.01900	-32.01929	10	-1.09
1372.50000	1387.65100	-15.15076	11	-.52
1369.70000	1377.46000	-7.76025	12	-.26
1374.60000	1374.74300	-.14282	13	.00
1436.50000	1379.49800	57.00171	14	1.95
1399.30000	1439.57300	-40.27332	15	-1.37
1421.90000	1425.21000	-3.30969	19	-.11
1422.70000	1425.40400	-2.70398	20	-.09
1427.00000	1426.18000	.81982	21	.03
1436.00000	1430.35400	5.64648	22	.19
1440.20000	1439.08800	1.11182	23	.04
1450.90000	1443.16400	7.73572	24	.26
1463.00000	1453.54900	9.45105	25	.32
1473.90000	1465.29200	8.60791	26	.29
1506.90000	1475.87100	31.02917	27	1.06
1526.00000	1507.89800	18.10205	28	.62

1517.30000	1526.43500	-9.13477	29	-.31
1491.70000	1517.99100	-26.29138	30	-.90
1429.50000	1493.14600	-63.64600	31	-2.17
1402.30000	1432.78000	-30.47974	32	-1.04
1383.00000	1406.38200	-23.38171	33	-.80
1393.90000	1387.65100	6.24927	34	.21
1398.10000	1398.22900	-.12939	35	.00
1488.00000	1402.30600	85.69446	36	2.93
1555.60000	1489.55500	66.04492	37	2.25
1525.20000	1555.16200	-29.96216	38	-1.02
1614.30000	1525.65800	88.64172	39	3.03
1613.20000	1612.13200	1.06836	40	.04
1579.00000	1611.06400	-32.06384	41	-1.09
1358.90000	1577.87200	-218.97220	42	-7.48
1411.80000	1364.26100	47.53882	43	1.62
1479.00000	1415.60200	63.39832	44	2.16
1508.80000	1480.82000	27.97961	45	.96
1461.70000	1509.74200	-48.04199	46	-1.64
1474.80000	1464.03000	10.76965	47	.37
1460.10000	1476.74400	-16.64429	48	-.57
1490.50000	1462.47800	28.02234	49	.96
1448.00000	1491.98100	-43.98145	50	-1.50
1422.50000	1450.73400	-28.23438	51	-.96
1412.70000	1425.98600	-13.28625	52	-.45
1467.30000	1416.47500	50.82495	53	1.74
1497.20000	1469.46500	27.73450	54	.95
1506.30000	1498.48400	7.81628	55	.27
1547.30000	1507.31600	39.98438	56	1.36
1429.10000	1547.10700	-118.00680	57	-4.03
1512.30000	1432.39200	79.90845	58	2.73
1530.20000	1513.13900	17.06128	59	.58
15[illegible]9.80000	1530.51100	-.71082	60	-.02
1516.50000	1530.12300	-13.62280	61	-.47
1422.40000	1517.21500	-94.81482	62	-3.24
1495.40000	1425.88900	69.51086	63	2.37
1500.60000	1496.73700	3.86304	64	.13
1488.60000	1501.78400	-13.18359	65	-.45
1516.30000	1490.13700	26.16272	66	.89
1512.50000	1517.02100	-4.52075	67	-.15
1505.00000	1513.33300	-8.33276	68	-.28
1497.70000	1506.05400	-8.35400	69	-.29
1436.10000	1498.96900	-62.86914	70	-2.15
1457.20000	1439.18500	18.01477	71	.61
1556.60000	1459.66300	96.93689	72	3.31
1546.60000	1556.13300	-9.53259	73	-.33
1508.90000	1546.42700	-37.52734	74	-1.28
1478.50000	1509.83900	-31.33899	75	-1.07
1481.00000	1480.33500	.66479	76	.02
1462.80000	1482.76100	-19.96143	77	-.68
1520.50000	1465.09800	55.40186	78	1.89
1532.20000	1521.09700	11.10303	79	.38
1500.10000	1532.45200	-32.35193	80	-1.10

1493.60000	1501.29800	-7.69836	81	-.26
1481.50000	1494.99000	-13.48999	82	-.46
1501.50000	1483.24700	18.25330	83	.62
1521.40000	1502.65700	18.74292	84	.64
1467.00000	1521.97000	-54.97046	85	-1.88
1466.10000	1469.17400	-3.07422	86	-.10
1480.60000	1468.30100	12.29919	87	.42
1544.10000	1482.37300	61.72668	88	2.11
1543.50000	1544.00100	-.50110	89	-.02
1539.90000	1543.41900	-3.51880	90	-.12
1490.50000	1539.92500	-49.42505	91	-1.69
1495.50000	1491.98100	3.51855	92	.12
1560.10000	1496.83400	63.26599	93	2.16
1559.40000	1559.52900	-.12939	94	.00
1548.50000	1558.85000	-10.35010	95	-.35
1584.70000	1548.27100	36.42847	96	1.24
1608.60000	1583.40400	25.19580	97	.86
1624.50000	1606.59900	17.90051	98	.61
1535.00000	1525.07600	9.92395	101	.34
1521.40000	1535.16900	-13.76941	102	-.47
1513.20000	1521.97000	-8.77051	103	-.30
1569.70000	1514.01200	55.68787	104	1.90
1555.60000	1568.84600	-13.24634	105	-.45
1512.30000	1555.16200	-42.86206	106	-1.46
1499.60000	1513.13900	-13.53870	107	-.46
1521.40000	1500.81300	20.58691	108	.70
1543.20000	1521.97000	21.22949	109	.72
1538.40000	1543.12800	-4.72766	110	-.16
1511.50000	1538.46900	-26.96924	111	-.92
1503.00000	1512.36200	-9.36230	112	-.32
3735.50000	1504.11300	2231.38700	113	76.17
5655.40000	3670.79400	1984.60500	114	67.75
4450.30000	5534.09200	-1083.79200	115	-37.00
2827.00000	4364.52100	-1537.52100	116	-52.49
1667.70000	2789.07900	-1121.37900	117	-38.28
1576.10000	1663.95700	-87.85718	118	-3.00
1587.00000	1575.05800	11.94226	119	.41
4287.10000	1585.63600	2701.46400	120	92.22
2072.00000	4206.13200	-2134.13200	121	-72.85
4156.30000	2056.33800	2099.96200	122	71.69
2898.60000	4079.18800	-1180.58800	123	-40.30
1513.80000	2858.56800	-1344.76800	124	-45.91
1805.60000	1514.59400	291.00550	125	9.93
3933.30000	1797.79200	2135.50800	126	72.90
2806.10000	3862.76300	-1056.66300	127	-36.07
1545.60000	2768.79500	-1223.19500	128	-41.76
1516.70000	1545.45700	-28.75696	129	-.98
1521.50000	1517.40900	4.09106	130	.14
1504.50000	1522.06700	-17.56738	131	-.60
1627.50000	1505.56900	121.93140	132	4.16
1475.00000	1624.94200	-149.94240	133	-5.12
1460.00000	1476.93800	-16.93835	134	-.58

1463.40000	1462.38100	1.01941	135	.03
1465.30000	1465.68000	-.38037	136	-.01
1493.20000	1467.52400	25.67554	137	.88
1472.70000	1494.60200	-21.90173	138	-.75
1493.70000	1474.70600	18.99377	139	.65
1482.70000	1495.08700	-12.38708	140	-.42
2023.80000	1484.41100	539.38880	141	18.41
1623.40000	2009.55900	-386.15870	142	-13.18
1527.90000	1620.96300	-93.06323	143	-3.18
1515.40000	1528.27900	-12.87878	144	-.44
1485.10000	1516.14700	-31.04736	145	-1.06
1527.80000	1486.74100	41.05945	146	1.40
1527.00000	1528.18200	-1.18176	147	-.04
1520.20000	1527.40500	-7.20532	148	-.25
1510.00000	1520.80600	-10.80566	149	-.37
1484.70000	1510.90600	-26.20654	150	-.89
1478.80000	1486.35200	-7.55225	151	-.26
1503.80000	1480.62600	23.17371	152	.79
1516.00000	1504.88900	11.11072	153	.38
1495.50000	1516.73000	-21.22961	154	-.72
1489.60000	1496.83400	-7.23401	155	-.25
1488.20000	1491.10800	-2.90796	156	-.10
1480.60000	1489.74900	-9.14917	157	-.31
1463.10000	1482.37300	-19.27332	158	-.66
1506.50000	1508.09200	-1.59192	161	-.05
1522.60000	1507.51000	15.09033	162	.52
1534.80000	1523.13500	11.66504	163	.40
1488.00000	1534.97500	-46.97534	164	-1.60
1483.00000	1489.55500	-6.55505	165	-.22

DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

--- L E A S T   M E D I A N   O F   S Q U A R E S ---

```
STAND. RESIDUAL I-+----+----+----+----+----+----+----   -------+----+----+-I
                I                                                     I
      .9222E+02 +   1                                                 +
                I                                                     I
                I  1                                                  I
                I      1  1                                           I
                I                            1                        I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                                                     +
                I  1                                                  I
                I                                                     I
                I  1                                                  I
                I 23                                                  I
            2.5 +++++++++++++++++++++++++++++++++++++++++++++++++++++++
            0.0 I-***-------------------------------------------------I
           -2.5 I+++++++++++++++++++++++++++++++++++++++++++++++++++++I
                I  123                                                I
                I        1                                            I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                  1           1                   1  +
                I                 1               1                   I
                I                  1                                  I
                I                                                     I
                I                                    1                I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
     -.7285E+02 +                                   1                 +
                I                                                     I
                I-+----+----+----+----+----+----+----+----+----+----+-I
          .1364E+04                                                .5534E+04

                                                ESTIMATED     KVALUE
```

```
               DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

                   --- L E A S T   M E D I A N   O F   S Q U A R E S ---
STAND. RESIDUAL
             I-+----+----+----+----+----+----+----+----+----+----+-I
             I                                                     I
   .9222E+02 +                                     1               +
             I                                                     I
             I                                   1                 I
             I                                     1 1             I
             I                                   1                 I
             +                                                     +
             I                                                     I
             I                                                     I
             I                                                     I
             I                                                     I
             +                                                     +
             I                                                     I
             I                                                     I
             I                                                     I
             I                                                     I
             +                                                     +
             I                                           1         I
             I                                                     I
             I                                      1              I
             I           11     1   1                 1            I
         2.5 +++++++++++++++++++++++++++++++++++++++++++++++++++++++
         0.0 I-433412343332234332233333433422343-1---323223343131-I
        -2.5 I+++++++++++++++++++++++++++++++++++++++++' ++++++++++I
             I             1    11                1    1  1        I
             I                                           1         I
             +                                                     +
             I                                                     I
             I                                                     I
             I                                                     I
             I                                                     I
             +                                   11  1             +
             I                                      11             I
             I                                      1              I
             I                                                     I
             I                                    1                I
             +                                                     +
             I                                                     I
             I                                                     I
             I                                                     I
             I                                                     I
  -.7285E+02 +                                     1               +
             I                                                     I
             I-+----+----+----+----+----+----+----+----+----+----+-I
              1                                                 165
                                                INDEX OF THE OBSERVATION
```

****

REWEIGHTED LEAST SQUARES BASED ON THE LMS

****

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE	P - VALUE
LAG1 VALUE	.84784	.04583	18.49899	.00000
CONSTANT	226.43360	68.25236	3.31759	.00118

WEIGHTED SUM OF SQUARES = 99323.21000

DEGREES OF FREEDOM = 129

SCALE ESTIMATE = 27.74793

VARIANCE - COVARIANCE MATRIX =

.2101D-02
-.3126D+01 .4658D+04

COEFFICIENT OF DETERMINATION (R SQUARED) = .72624

THE F-VALUE = 342.213 (WITH 1 AND 129 DF) P - VALUE = .00000

THERE ARE 131 POINTS WITH NON-ZERO WEIGHT.

AVERAGE WEIGHT = .82911

OBSERVED KVALUE	ESTIMATED KVALUE	RESIDUAL	NO	RES/SC	WEIGHT
1528.80000	1535.06800	-6.26758	1	-.23	1.0
1473.80000	1522.60400	-48.80432	2	-1.76	1.0
1422.70000	1475.97400	-53.27356	3	-1.92	1.0
1389.90000	1432.64900	-42.74902	4	-1.54	1.0
1382.00000	1404.84000	-22.84009	5	-.82	1.0
1406.20000	1398.14200	8.05786	6	.29	1.0
1384.50000	1418.66000	-34.15967	7	-1.23	1.0
1445.80000	1400.26200	45.53833	8	1.64	1.0
1411.20000	1452.23400	-41.03418	9	-1.48	1.0
1383.00000	1422.89900	-39.89893	10	-1.44	1.0
1372.50000	1398.99000	-26.48999	11	-.95	1.0
1369.70000	1390.08800	-20.38770	12	-.73	1.0
1374.60000	1387.71400	-13.11377	13	-.47	1.0
1436.50000	1391.86800	44.63184	14	1.61	1.0
1399.30000	1444.34900	-45.04907	15	-1.62	1.0

1421.90000	1431.80100	-9.90112	19	-.36	1.0
1422.70000	1431.97100	-9.27087	20	-.33	1.0
1427.00000	1432.64900	-5.64905	21	-.20	1.0
1436.00000	1436.29500	-.29480	22	-.01	1.0
1440.20000	1443.92500	-3.72534	23	-.13	1.0
1450.90000	1447.48600	3.41394	24	.12	1.0
1463.00000	1456.55800	6.44202	25	.23	1.0
1473.90000	1466.81700	7.08325	26	.26	1.0
1506.90000	1476.05800	30.84180	27	1.11	1.0
1526.00000	1504.03700	21.96313	28	.79	1.0
1517.30000	1520.23000	-2.93042	29	-.11	1.0
1491.70000	1512.85400	-21.15442	30	-.76	1.0
1429.50000	1491.15000	-61.64966	31	-2.22	1.0
1402.30000	1438.41400	-36.11426	32	-1.30	1.0
1383.00000	1415.35300	-32.35327	33	-1.17	1.0
1393.90000	1398.99000	-5.08997	34	-.18	1.0
1398.10000	1408.23100	-10.13147	35	-.37	1.0
1488.00000	1411.79200	76.20776	36	2.75	.0
1555.60000	1488.01300	67.58728	37	2.44	1.0
1525.20000	1545.32600	-20.12646	38	-.73	1.0
1614.30000	1519.55200	94.74792	39	3.41	.0
1613.20000	1595.09400	18.10559	40	.65	1.0
1579.00000	1594.16200	-15.16162	41	-.55	1.0
1358.90000	1565.16600	-206.26570	42	-7.43	.0
1411.80000	1378.55700	33.24292	43	1.20	1.0
1479.00000	1423.40800	55.59229	44	2.00	1.0
1508.80000	1480.38200	28.41785	45	1.02	1.0
1461.70000	1505.64800	-43.94775	46	-1.58	1.0
1474.80000	1465.71500	9.08545	47	.33	1.0
1460.10000	1476.82100	-16.72131	48	-.60	1.0
1490.50000	1464.35800	26.14197	49	.94	1.0
1448.00000	1490.13200	-42.13232	50	-1.52	1.0
1422.50000	1454.09900	-31.59924	51	-1.14	1.0
1412.70000	1432.47900	-19.77954	52	-.71	1.0
1467.30000	1424.17100	43.12939	53	1.55	1.0
1497.20000	1470.46300	26.73743	54	.96	1.0
1506.30000	1495.81300	10.48730	55	.38	1.0
1547.30000	1503.52800	43.77197	56	1.58	1.0
1429.10000	1538.28900	-109.18950	57	-3.94	.0
1512.30000	1438.07500	74.22485	58	2.67	.0
1530.20000	1508.61500	21.58484	59	.78	1.0
1529.80000	1523.79100	6.00879	60	.22	1.0
1516.50000	1523.45200	-6.95227	61	-.25	1.0
1422.40000	1512.17600	-89.77600	62	-3.24	.0
1495.40000	1432.39500	63.00525	63	2.27	1.0
1500.60000	1494.28700	6.31323	64	.23	1.0
1488.60000	1498.69500	-10.09546	65	-.36	1.0
1516.30000	1488.52100	27.77869	66	1.00	1.0
1512.50000	1512.00600	.49353	67	.02	1.0
1505.00000	1508.78500	-3.78467	68	-.14	1.0
1497.70000	1502.42600	-4.72595	69	-.17	1.0
1436.10000	1496.23700	-60.13672	70	-2.17	1.0

1457.20000	1444.01000	13.18994	71	.48	1.0
1556.60000	1461.89900	94.70068	72	3.41	.0
1546.60000	1546.17400	.42578	73	.02	1.0
1508.90000	1537.69600	-28.79578	74	-1.04	1.0
1478.50000	1505.73200	-27.23242	75	-.98	1.0
1481.00000	1479.95800	1.04175	76	.04	1.0
1462.80000	1482.07800	-19.27783	77	-.69	1.0
1520.50000	1466.64700	53.85266	78	1.94	1.0
1532.20000	1515.56700	16.63257	79	.60	1.0
1500.10000	1525.48700	-25.38696	80	-.91	1.0
1493.60000	1498.27100	-4.67151	81	-.17	1.0
1481.50000	1492.76000	-11.26050	82	-.41	1.0
1501.50000	1482.50200	18.99817	83	.68	1.0
1521.40000	1499.45800	21.94153	84	.79	1.0
1467.00000	1516.33000	-49.33044	85	-1.78	1.0
1466.10000	1470.20800	-4.10815	86	-.15	1.0
1480.60000	1469.44500	11.15491	87	.40	1.0
1544.10000	1481.73900	62.36133	88	2.25	1.0
1543.50000	1535.57600	7.92371	89	.29	1.0
1539.90000	1535.06800	4.83240	90	.17	1.0
1490.50000	1532.01500	-41.51538	91	-1.50	1.0
1495.50000	1490.13200	5.36768	92	.19	1.0
1560.10000	1494.37100	65.72852	93	2.37	1.0
1559.40000	1549.14200	10.25842	94	.37	1.0
1548.50000	1548.54800	-.04822	95	.00	1.0
1584.70000	1539.30700	45.39319	96	1.64	1.0
1608.60000	1569.99800	38.60168	97	1.39	1.0
1624.50000	1590.26200	34.23840	98	1.23	1.0
1535.00000	1519.04300	15.95654	101	.58	1.0
1521.40000	1527.86100	-6.46094	102	-.23	1.0
1513.20000	1516.33000	-3.13049	103	-.11	1.0
1569.70000	1509.37800	60.32190	104	2.17	1.0
1555.60000	1557.28100	-1.68079	105	-.06	1.0
1512.30000	1545.32600	-33.02637	106	-1.19	1.0
1499.60000	1508.61500	-9.01514	107	-.32	1.0
1521.40000	1497.84800	23.55249	108	.85	1.0
1543.20000	1516.33000	26.86951	109	.97	1.0
1538.40000	1534.81300	3.58679	110	.13	1.0
1511.50000	1530.74400	-19.24365	111	-.69	1.0
1503.00000	1507.93700	-4.93689	112	-.18	1.0
3735.50000	1500.73000	2234.77000	113	80.54	.0
5655.40000	3393.52300	2261.87700	114	81.52	.0
4450.30000	5021.28200	-570.98190	115	-20.58	.0
2827.00000	3999.55500	-1172.55500	116	-42.26	.0
1667.70000	2623.26400	-955.56450	117	-34.44	.0
1576.10000	1640.36900	-64.26868	118	-2.32	.0
1587.00000	1562.70700	24.29297	119	.88	1.0
4287.10000	1571.94800	2715.15200	120	97.85	.0
2072.00000	3861.18900	-1789.18900	121	-64.48	.0
4156.30000	1983.14900	2173.15100	122	78.32	.0
2898.60000	3750.29200	-851.69170	123	-30.69	.0
1513.80000	2683.96900	-1170.16900	124	-42.17	.0

1805.60000	1509.88700	295.71310	125	10.66	.0
3933.30000	1757.28500	2176.01500	126	78.42	.0
2806.10000	3561.22500	-755.12450	127	-27.21	.0
1545.60000	2605.54500	-1059.94500	128	-38.20	.0
1516.70000	1536.84800	-20.14807	129	-.73	1.0
1521.50000	1512.34600	9.15442	130	.33	1.0
1504.50000	1516.41500	-11.91516	131	-.43	1.0
1627.50000	1502.00200	125.49800	132	4.52	.0
1475.00000	1606.28600	-131.28580	133	-4.73	.0
1460.00000	1476.99100	-16.99084	134	-.61	1.0
1463.40000	1464.27300	-.87329	135	-.03	1.0
1465.30000	1467.15600	-1.85596	136	-.07	1.0
1493.20000	1468.76700	24.43311	137	.88	1.0
1472.70000	1492.42100	-19.72144	138	-.71	1.0
1493.70000	1475.04100	18.65918	139	.67	1.0
1482.70000	1492.84500	-10.14539	140	-.37	1.0
2023.80000	1483.51900	540.28090	141	19.47	.0
1623.40000	1942.28300	-318.88290	142	-11.49	.0
1527.90000	1602.81000	-74.90967	143	-2.70	.0
1515.40000	1521.84100	-6.44128	144	-.23	1.0
1485.10000	1511.24300	-26.14343	145	-.94	1.0
1527.80000	1485.55400	42.24609	146	1.52	1.0
1527.00000	1521.75700	5.24341	147	.19	1.0
1520.20000	1521.07800	-.87830	148	-.03	1.0
1510.00000	1515.31300	-5.31299	149	-.19	1.0
1484.70000	1506.66500	-21.96509	150	-.79	1.0
1478.80000	1485.21500	-6.41479	151	-.23	1.0
1503.80000	1480.21300	23.58740	152	.85	1.0
1516.00000	1501.40900	14.59143	153	.53	1.0
1495.50000	1511.75200	-16.25208	154	-.59	1.0
1489.60000	1494.37100	-4.77148	155	-.17	1.0
1488.20000	1489.36900	-1.16931	156	-.04	1.0
1480.60000	1488.18200	-7.58228	157	-.27	1.0
1463.10000	1481.73900	-18.63867	158	-.67	1.0
1506.50000	1504.20600	2.29370	161	.08	1.0
1522.60000	1503.69800	18.90234	162	.68	1.0
1534.80000	1517.34800	17.45227	163	.63	1.0
1488.00000	1527.69100	-39.69141	164	-1.43	1.0
1483.00000	1488.01300	-5.01270	165	-.18	1.0

```
        DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

      --- R E W E I G H T E D   L S     ( B A S E D   O N   L M S ) ---

STAND. RESIDUAL I-+----+----+----+----+----+----+----+----+----+----+-I
                I                                                     I
     .9785E+02 +    1                                                 +
                I                                                     I
                I                                                     I
                I  1   1  1                  1                        I
                I                                                     I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
                +                                                     +
                I                                                     I
                I                                                     I
                I  1                                                  I
                I                                                     I
                +  1                                                  +
                I 23                                                  I
           2.5 I++1++++++++++++++++++++++++++++++++++++++++++++++++++++I
           0.0 I-***1-------------------------------------------------I
          -2.5 I++1+1++++++++++++++++++++++++++++++++++++++++++++++++++I
                +   21   1                                            +
                I                                                     I
                I                                                     I
                I                                                   1 I
                I                             1                       I
                +                                1                    +
                I                  1                                  I
                I                 1                                   I
                I                  1                 1                I
                I                                                     I
                +                                                     +
                I                                                     I
                I                                                     I
                I                                                     I
                I                                                     I
    -.6448E+02 +                                   1                  +
                I                                                     I
                I-+----+----+----+----+----+----+----+----+----+----+-I
          .1379E+04                                             .5021E+04

                                          ESTIMATED     KVALUE
```

DAYS 91251-91365 OF 889 YR 1991 USING KVALUE AND LAG ONE

```
   --- R E W E I G H T E D   L S     ( B A S E D   O N   L M S ) ---
STAND. RESIDUAL
           I-+----+----+----+----+----+----+----+----+----+----+-I
           I                                                    I
 .9785E+02 +                                     1              +
           I                                                    I
           I                                                    I
           I                                   2 1 1            I
           I                                                    I
           +                                                    +
           I                                                    I
           I                                                    I
           I                                                    I
           I                                                    I
           +                                                    +
           I                                                    I
           I                                                    I
           I                                                    I
           I                                                    I
           +                                                    +
           I                                                    I
           I                                                    I
           I                                          1         I
           I                                                    I
           +                                     1              +
           I          11      1   1                1            I
       2.5 I+++++++++++1+++++++++++++++++++++++++++++++++++++++++I
       0.0 I-433412343322234332233333433422343-2---323223343131-I
      -2.5 I+++++++++++++++++++1+++++++++++++++++++++++1++++++++I
           +             1    1                      1 1        +
           I                                                    I
           I                                                    I
           I                                   1                I
           I                                       1            I
           +                                      1             +
           I                                    1               I
           I                                       1            I
           I                                    1 1             I
           I                                                    I
           +                                                    +
           I                                                    I
           I                                                    I
           I                                                    I
           I                                                    I
-.6448E+02 +                                    1               +
           I                                                    I
           I-+----+----+----+----+----+----+----+----+----+----+-I
             1                                             165

                                 INDEX OF THE OBSERVATION
```

Appendix E: Corrolagrams

This appendix presents the correlograms, the ACF and PACF plots, used in the results chapter. These plots were used to aid in the determination of the order of autoregression appropriate for the AR(1)-RLS model in the Results chapter. These correlograms were created using S-Plus Statistical Software.

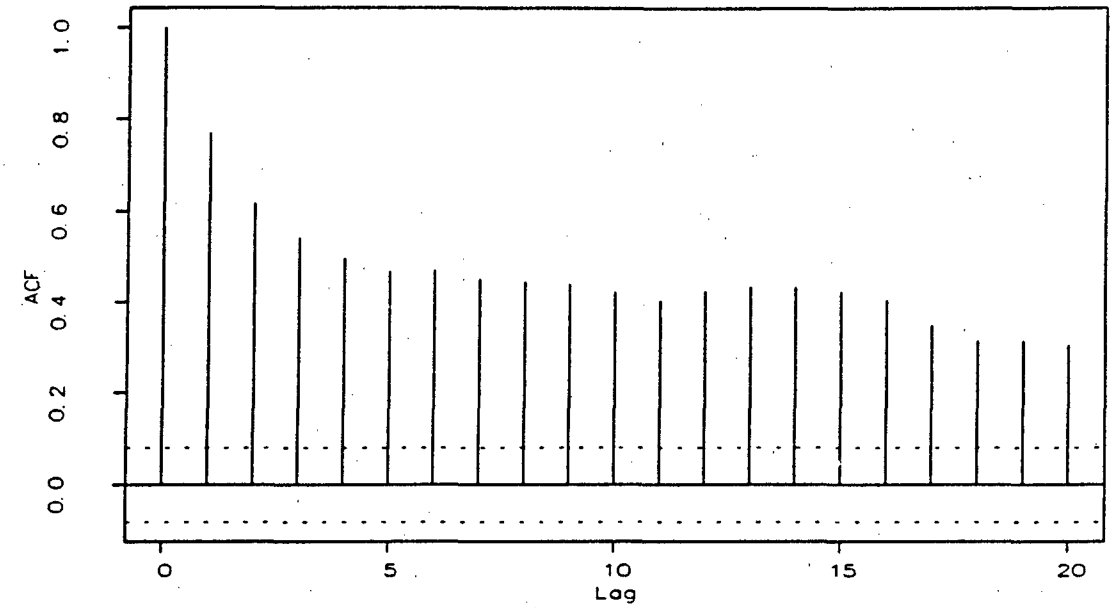

Figure 19. ACF Plot For Site 852

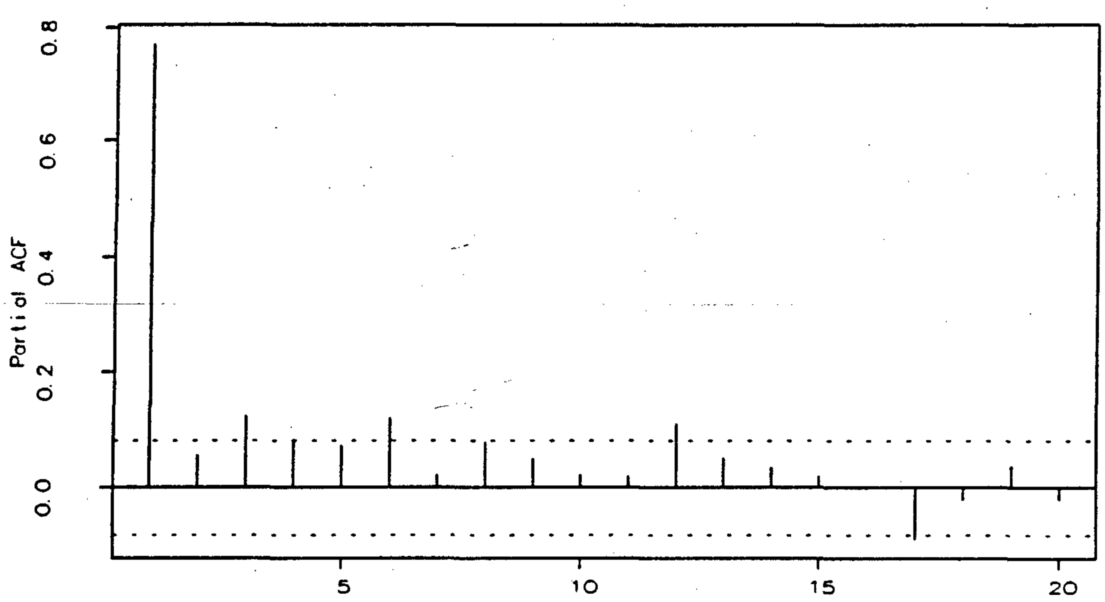

Figure 20. PACF Plot For Site 852

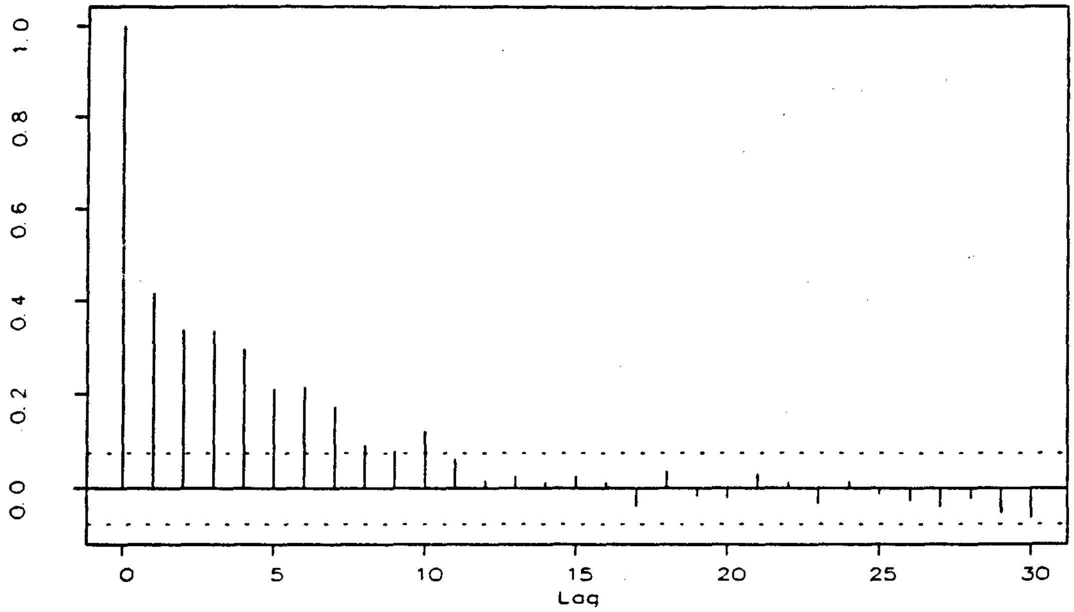

Figure 21. ACF Plot for Site 858.

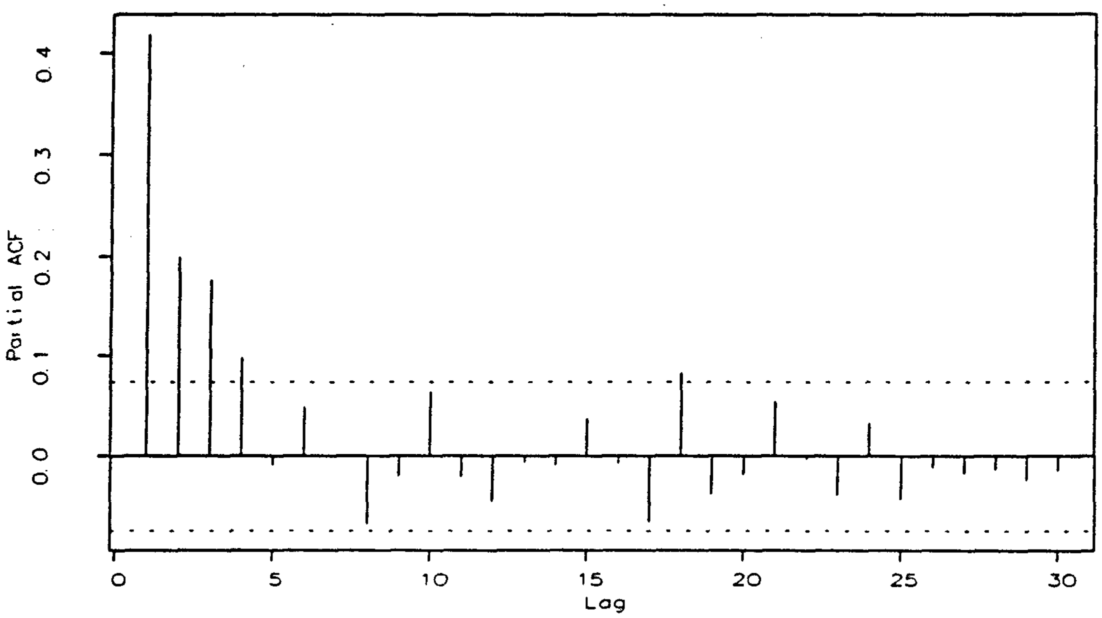

Figure 22. PACF Plot for Site 858.

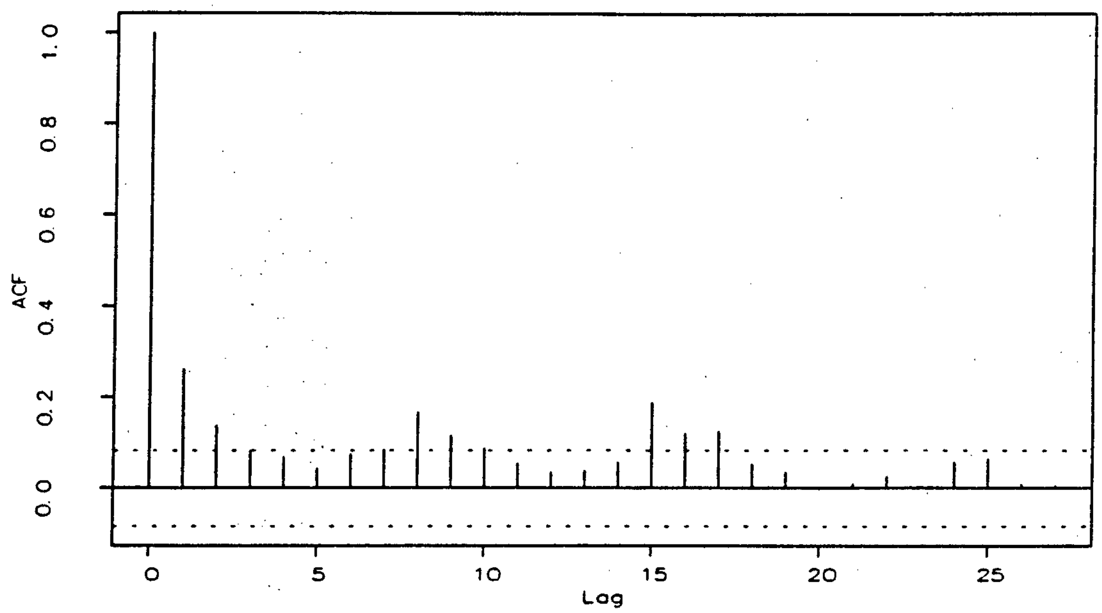

Figure 23. ACF Plot for Site 889.

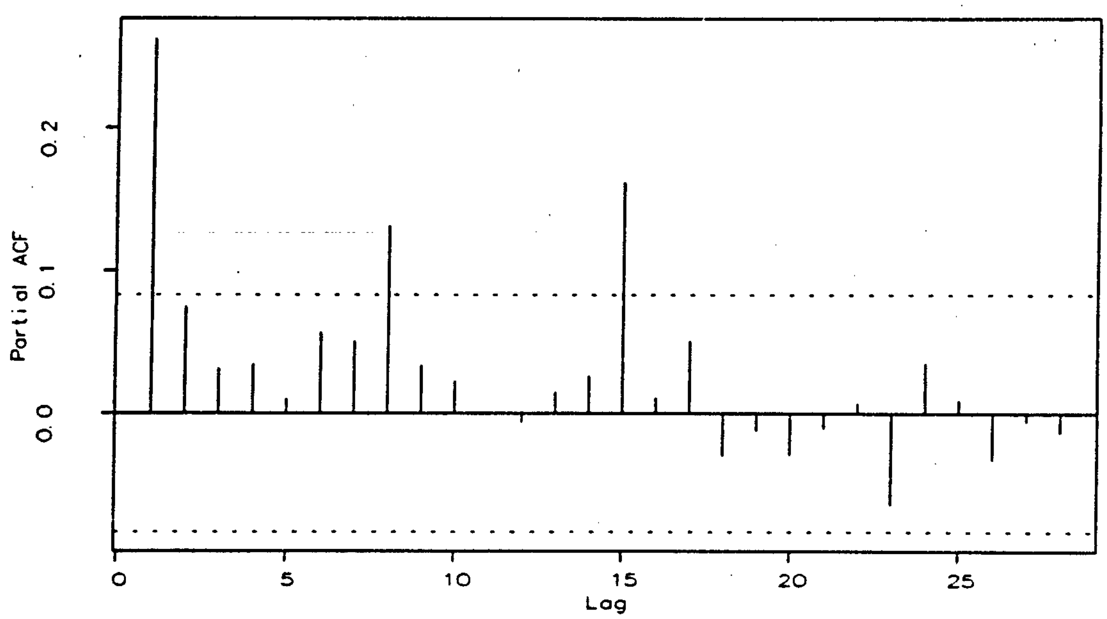

Figure 24. PACF Plot for Site 889.

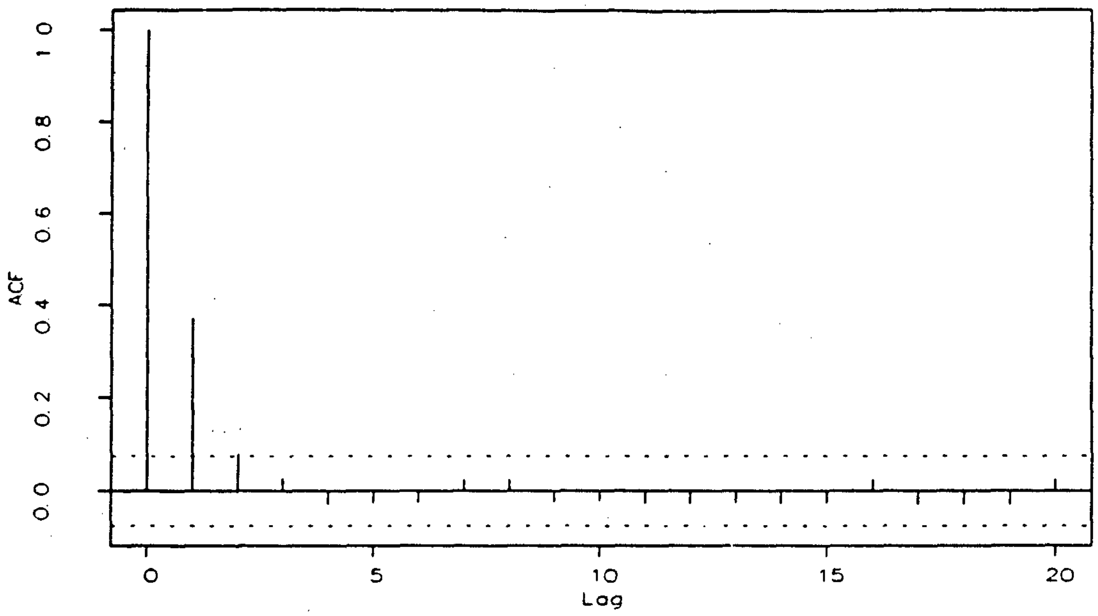

Figure 25. ACF Plot for Site 981.

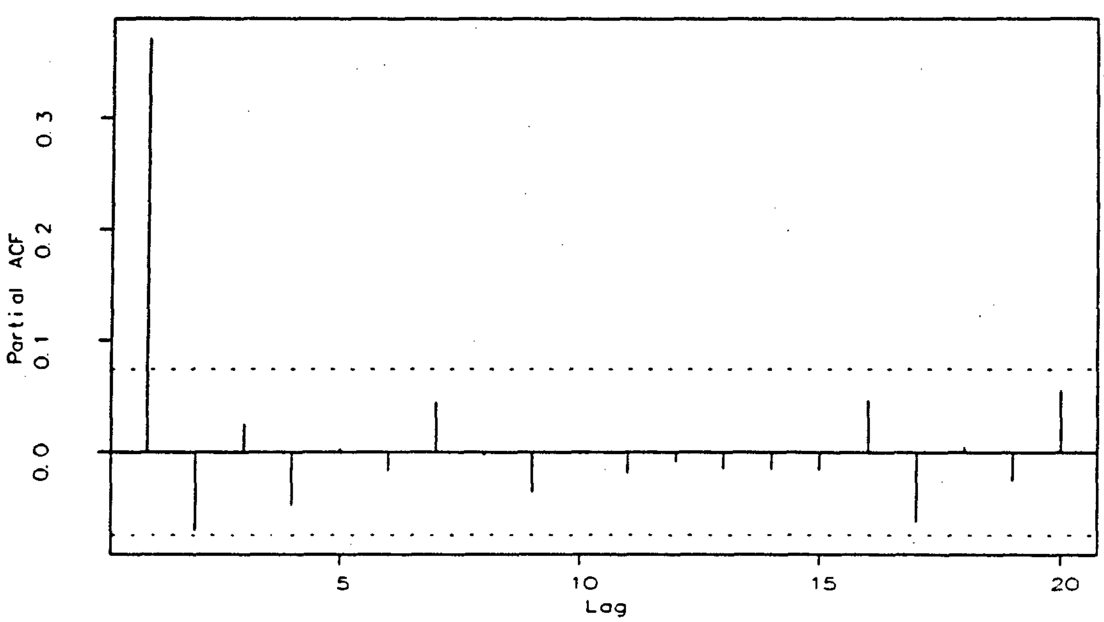

Figure 26. PACF Plot for Site 981.

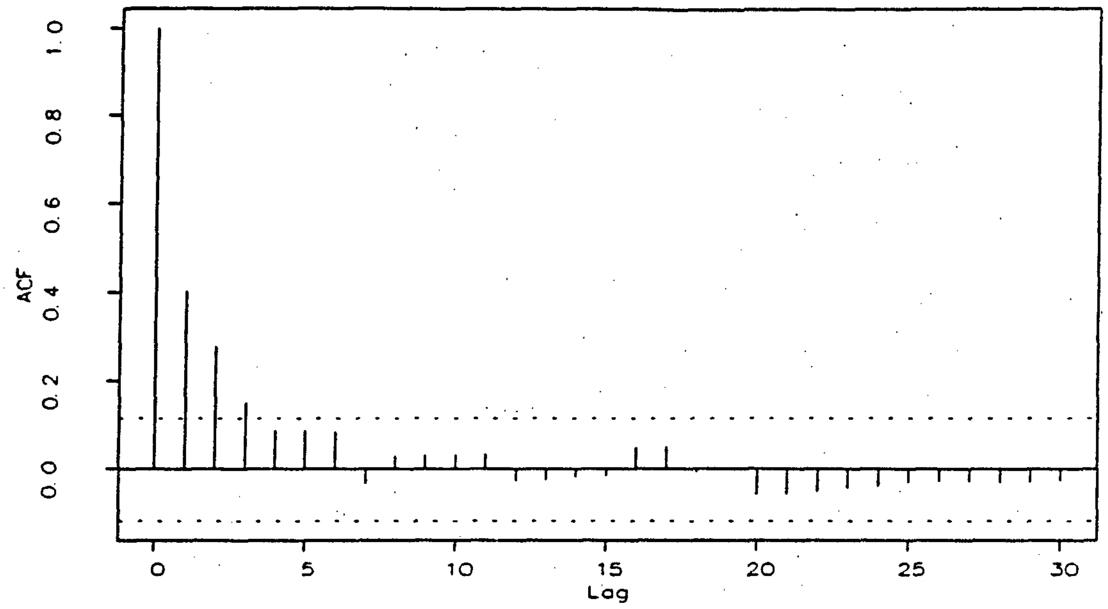

Figure 27. ACF Plot for Site 996

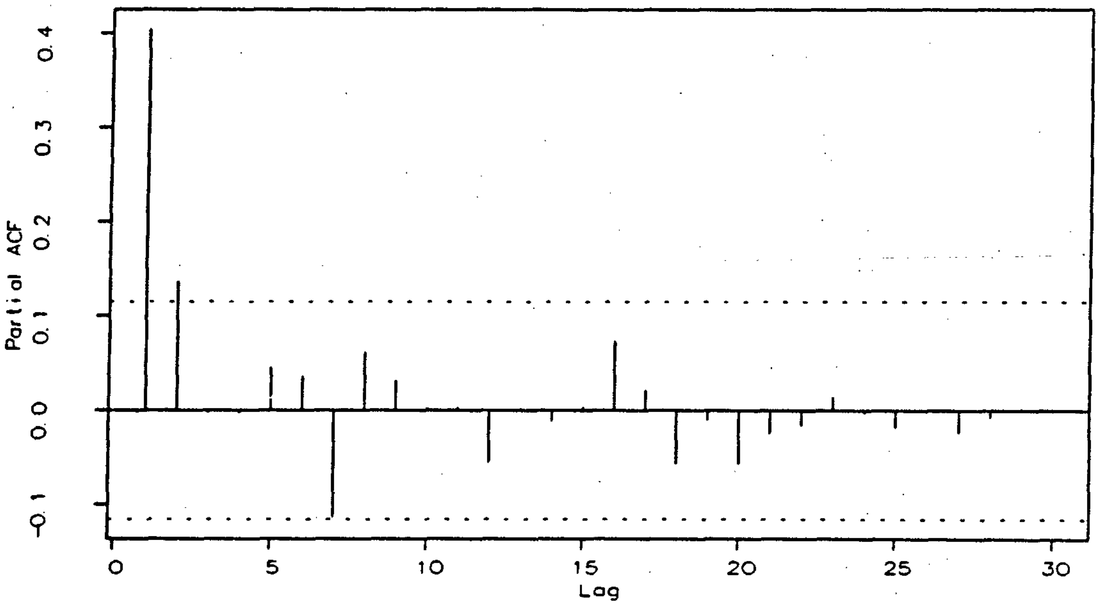

Figure 28. PACF Plot for Site 996

Bibliography

1. Abraham, Bovas, and Johannes Ledolter. *Statistical Methods for Forecasting*. New York: John Wiley & Sons, Inc. 1983.

2. Anscombe, F.J. "Graphs in Statistical Analysis." *American Statistician*, 27: 17-21 (February, 1973).

3. Bond, Walter P., and J. Michael Sonnier. "Significant Data Identification." Unpublished Report No. ARS-SSR-86-08. ENSCO, Inc. Indian Harbour Beach, FL. January 1986.

4. Box, George P. and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control* (Revised Edition). London: Holden-Day, 1976.

5. Brockwell, Peter J. and Richard A. Davis. *Time Series: Theory and Methods* (Second Edition). New York: Springer-Verlag, 1991.

6. Chatfield, Christopher. *The Analysis of Time Series* (Third Edition). New York: Chapman and Hall, Ltd., 1984.

7. Currie, Dr. Lloyd A, Telephone interviews. National Institute of Standards and Technology, United States Department of Commerce, Gathersburg MD, 1 August through 20 November, 1992.

8. Currie, Dr. Lloyd A, "Pseudo-Code for Plume Detection." Unpublished. 18 July 1991.

9. Diggle, Peter J. *Time Series: A Biostatistical Introduction.* New York: Oxford University Press, 1990.

10. ENSCO Inc. *Three Collections of FORTRAN Routines: LOWESS, Loess, and STL.* Unpublished. 5 February 1989.

11. Hoaglin, David C. and others. *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons, 1983.

12. Hoff, John C. *A Practical Guide to BOX-JENKINS Forecasting.* Belmont CA: Lifetime Learning Publications, 1983.

13. Jenkins, Gwilym M. and Donald G. Watts. *Spectral Analysis and it's Applications.* San Francisco: Holden-Day, 1968.

14. Makridakis, Spyros and others. *Forecasting: Methods and Applications* (Second Edition). New York: John Wiley & Sons, 1983.

15. Martin, R. D. and V.J. Yohai. "Robustness in Time Series and Estimating ARMA Models." *Handbook of Statistics* 5, edited by E.J. Hannan et al. New York: North-Holland Publishing Company, 1985.

16. McCleary, Richard and Others. *Applied Time Series Analysis for the Social Sciences*. London: Sage Publications, 1980.

17. Mykytka, Edward F. Professor, Air Force Institute of Technology. Personal interview

18. Rousseeuw, Peter J. "Least Median of Squares Regression", *Journal of the American Statistical Association*, 79: 871-880 (December 1984)

19. Rousseeuw, Peter J. Professor. Personal Correspondence. Vesaliuslaan 24, B-2650 Edegem Belgium, 24 November 1992.

20. Rousseeuw, Peter J. and Annick M. Leroy. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, Inc. 1987.

21. Ryan, Thomas P. *Statistical Methods for Quality Improvement*. New York : John Wiley & Sons, Inc. 1989.

22. Schaefer, Robert L. and Elizabeth Farber, *The Student Edition of Minitab* (Version 8). Reading, MA: Addison-Wesley Publishing Company, Inc. 1992.

23. Tinsley, Capt. Russell, Nuclear Systems Analyst. Personal interviews. Nuclear Reactors Division, Nuclear Technology Directorate, Headquarters, Air Force Technical Applications Center, Patrick AFB FL, 1 July through 15 October 1992.

24. Wolfram, Stephen. *Mathematica: A System for Doing Mathematics by Computer* (Second Edition). Redwood City, CA: Addison-Wesley Publishing Company, Inc. 1991.

Vita

Captain Keri L. Robinson was born on 20 June 1958 in Tampa, Florida. He is the son of William and Karol Robinson of Stone Mountain, Georgia. He graduated from high school in Clarkston, Georgia in 1976. In 1977 he enlisted in the United States Air Force and served as a Korean and German translator until 1983 when he was accepted into the Airman's Education and Commissioning Program. He attended The Georgia Institute of Technology, Atlanta, Georgia, from which he received the degree of Bachelor of Nuclear Engineering in March 1986. Upon graduation, he received his commission from Officer Training School in July 1986. He served two years at the Oklahoma City Air Logistics Center as a Nuclear Survivability Officer for Strategic Air Command's air breathing platforms at Tinker AFB, Oklahoma. He was then assigned to the Air Force Technical Application Center, Patrick AFB, Florida where he worked as a Nuclear Systems Analyst until entering the School of Engineering, Air Force Institute of Technology, in August 1991. Captain Robinson is married to the former Linda S. Reid of Stone Mountain, Georgia. They have two children, Alicia and Kristopher.

Permanent Address: 4414 Hwy 138 SW
Stockbridge, Georgia

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE

Identification of Significant Outliers in Time Series Data

5. FUNDING NUMBERS

6. AUTHOR(S)

Keri L. Robinson, Capt, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology, WPAFB OH 45433-6853

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GNE/ENP/93M-7

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Technical Applications Center/TNR
Patrick AFB, FL 32925
POC: Capt Donald Culp DSN 894-6365

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This thesis examines the feasibility of using least median of squares (LMS) procedure applied to a reweighted least squares (RLS) autoregression model to identify significant outliers in time series data. The time series were analyzed for data points that were outliers. In order to perform detailed analysis on an outlier, the analyst must be able to determine that an outlier data point is significantly different from normally distributed data. This thesis examines a new method for identifying these outliers. Data from the field were characterized and fit with time series models using an autoregressive reweighted least squares routine (ARRLS) derived from the LMS methodology. Various orders of autoregression were applied to the ARRLS method to determine an appropriate order for the model; resulting fit coefficients were tested for significance. Regression results from data taken at five sites are presented. By using an autoregressive order of one (AR(1)) applied to the ARRLS, this method significantly improved outlier detection in the time series data over the recursive removal without regression (RRR) method currently in use. In addition to identifying the outliers found by RRR, the AR(1)-RLS method routinely identified four times as many outliers as AFTAC's RRR method. The AR(1)-RLS method is recommended as a complimentary procedure to the RRR method currently used in identifying significant outliers. After sufficient operational experience is gained, AR(1)-RLS may supplant current schemes. Recommendations for improvements to the AR(1)-RLS method are offered.

14. SUBJECT TERMS

OUTLIER, LEAST SQUARES, AUTOREGRESSION, LEAST MEDIAN SQUARED RESIDUALS

15. NUMBER OF PAGES

146

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

END

FILMED

DATE: 4-93

DTIC